DISS. ETH NO. 24082

Bayesian techniques for inverse uncertainty quantification

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZÜRICH
(Dr. sc. ETH Zürich)

Presented by
JOSEPH BENJAMIN NAGEL
Diplom-Physiker, University of Bonn
Born on 27.09.1985
Citizen of Germany

Accepted on the recommendation of
Prof. Dr. Bruno Sudret, examiner
Prof. Dr. Costas Papadimitriou, co-examiner
Prof. Dr. Eleni Chatzi, co-examiner

2017

Abstract

Bayesian inference enables the fusion of heterogeneous information and the reduction of epistemic uncertainty for solving inverse problems. Physical models, expert knowledge and experimental data are statistically interpreted in order to learn about the unknown model parameters. The prior and the posterior probability distribution represent the uncertainty of the unknowns before and after the analysis. Bayes' theorem governs the update from the prior to the conditional posterior which reflects the achieved gain of information.

Characterizing the posterior distribution poses the main challenge in Bayesian data analysis. Since only very simple problems admit analytical solutions, most often one has to compute the posterior numerically. This formidable task is accomplished by means of Markov chain Monte Carlo techniques. The large number of forward model runs that is thereby required prohibits computational inference in many fields of application. In civil, mechanical and aerospace engineering this holds especially true.

The goal of this doctoral dissertation is to develop new approaches to the Bayesian probabilistic analysis in complex and realistic engineering applications. A unified framework for inverse problems under epistemic uncertainty and aleatory variability is elaborated to that end. Hamiltonian Monte Carlo is used as an efficient sampling algorithm that overcomes the associated computational difficulties. Moreover, completely novel methods for posterior computation are presented and investigated.

Zusammenfassung

Bayessche Inferenz erlaubt die Miteinbeziehung heterogener Informationen und die Reduktion epistemischer Unsicherheiten beim Lösen inverser Probleme. Physikalische Modelle, Expertenwissen und Messdaten werden statistisch ausgewertet, um auf die unbekannten Modellparameter zurückzuschließen. Die Prior- und die Posterior-Wahrscheinlichkeitsverteilung repräsentieren die Unsicherheit der Unbekannten vor und nach der Analyse. Der Übergang vom Prior in den bedingten Posterior erfolgt nach dem Lehrsatz von Bayes und spiegelt den erzielten Informationsgewinn wieder.

Das Charakterisieren der Posterior-Verteilung stellt die größte Herausforderung der Bayesschen Datenanalyse dar. Weil nur denkbar einfache Probleme analytische Lösungen besitzen, muss man den Posterior meistens numerisch berechnen. Markov-Ketten-Monte-Carlo-Verfahren werden zur Bewältigung dieser schwierigen Aufgabe eingesetzt. Die große Anzahl der dafür erforderlichen Vorwärtsmodellläufe verhindert die rechnergestützte Inferenz in vielen Anwendungsgebieten. Dies gilt insbesondere im Bauingenieurwesen und Maschinenbau sowie in der Luft- und Raumfahrttechnik.

Ziel dieser Doktorarbeit ist es, neue Methoden zur Bayesschen Wahrscheinlichkeitsanalyse in komplexen und realistischen Ingenieuranwendungen zu entwickeln. Ein einheitliches Rahmenkonzept für inverse Probleme unter epistemischer Unsicherheit und aleatorischer Variabilität wird zu diesem Zweck ausgearbeitet. Um den damit verbundenen Rechenaufwand zu verringern, wird ein effizienter Hamiltonscher Monte-Carlo-Algorithmus verwendet. Darüber hinaus werden völlig neuartige Ansätze zur Berechnung der Posterior-Wahrscheinlichkeitsverteilung vorgestellt und untersucht.

Table of contents

Chapter 1

Overview

1.1 Motivation

The increasing sophistication of computer simulations for predicting the behavior of physical systems necessitates the specification of a growing number of model parameters. This motivates the engagement in both forward and inverse uncertainty quantification. One can represent the degree as to which the model input parameters are not precisely known as a probability distribution. Either this may reflect a lack of knowledge about the true parameter value or a natural variability of the input realizations. The stochasticity in the model parameters then induces randomness in the model predictions, the quantification of which is the chief goal of uncertainty forward propagation.

While uncertainty propagation deals with the characterization of the model response for a given input distribution, inverse uncertainty quantification aims at the indirect determination of the actual distribution of the uncertain inputs with experimental measurements of the outputs. In Bayesian inverse problems the epistemic uncertainty of the constant but unknown forward model parameters is translated into the prior distribution and probabilistically updated. The resulting posterior distribution encodes the reduced level of epistemic uncertainty that remains after integrating the information yielded by the data. Point estimates of the parameters and predictive distributions of future outcomes can then be derived.

The Bayesian approach to inverse problems does not only allow for improving one's knowledge about the fixed yet unknown parameters and growing one's confidence in the predictions, it also measures the uncertainty in the model input estimation and output prediction. Therefore it gains advantage over deterministic solutions to inverse problems. A limitation of the approach is that it lacks the possibility to manage the aleatory uncertainty of genuinely random quantities that vary during the experimentation. Nuisance variables that merely complicate the analysis or aleatory variables whose distribution is of inferential interest are examples of such quantities. They are incorrectly treated as constants in current practice.

In addition to the unanswered question of how aleatory variability might be handled, another limiting factor of Bayesian inversion is the expense of computing the posterior distribution numerically. One of the very few serviceable tools for that purpose is Markov chain Monte Carlo sampling. This technique suffers from the absence of a clear convergence criterion and the autocorrelation of the obtained posterior samples. It demands an excessive number of serial forward model runs which may easily exceed the available computational budget. This prompts researchers and practitioners to implement advanced sampling algorithms and to find completely new solutions.

1.2 Contribution

The fact that aleatory variability is ignored, the need for more efficient sampling schemes and the lack of fundamental alternatives form obstacles to Bayesian inverse problems in complex applications. In this dissertation it is tried to overcome these difficulties. The core contributions are concisely summarized as follows.

1) A unifying framework for the management of aleatory uncertainty in inverse problems is developed.

2) Hamiltonian Monte Carlo is promoted for efficient posterior exploration in high-dimensional spaces.

3) Novel approaches for computational Bayesian inference and posterior characterization are proposed.

4) Complex inverse problems from structural, mechanical and also hydrological engineering are solved.

First, the developed framework in 1) allows one to master inverse problems in the presence of epistemic uncertainty and aleatory variability. Unknown parameters can be identified along with the distribution of aleatory variables. Second, in 2) the computational cost of sampling high-dimensional posteriors is drastically reduced by Hamiltonian Monte Carlo. This is a general-purpose Markov chain Monte Carlo method which proves especially beneficial to the previously devised framework for inversion under polymorphic uncertainty. Third, spectral Bayesian inference is proposed as a radically different technique for computing the posterior density in 3). It rests on spectral likelihood expansions and enables semi-analytic and sampling-free Bayesian inference. Another recently emerged method based on optimal transportation theory is also investigated and compared to spectral inference.

A wide range of practical engineering problems can be addressed with the new methodological developments. On the one hand, simple problems with simulated data serve for prototyping and benchmarking purposes. Bayesian inversion under multiple types of uncertainty and Hamiltonian Monte Carlo are both applied to the estimation of the material variability throughout an ensemble of structural elements. The inverse heat conduction problem posed by calibrating the thermal properties of a composite material with temperature measurements is used to demonstrate the newly devised schemes of inference. On the other hand, in 4) some more interesting problems involving real data and realistic models are solved. This includes the NASA Langley multidisciplinary uncertainty quantification challenge, the probabilistic assessment of structural masonry, and the Bayesian calibration of a hydrological urban drainage simulator.

The methodological progress achieved and the real problems solved are the main outcomes of this doctoral research work. They have led to four journal publications [1–4], an equal number of conference papers [5–8] and five other presentations [9–13]. The most important contributions [1–4, 8] are contained as individual chapters later on in the dissertation. Postprints of the finally accepted and already published articles [1–4] after scholarly peer review and before the copyediting and typesetting are provided. The conference paper [8] is supplemented with three additional graphics whose inclusion was originally prevented by the template and page limit.

1.2.1 Journal papers

[1] J. B. Nagel and B. Sudret. “Bayesian Multilevel Model Calibration for Inverse Problems Under Uncertainty with Perfect Data”. In: *Journal of Aerospace Information Systems* 12.1 (2015), pp. 97–113. DOI: 10.2514/1.I010264.

[2] J. B. Nagel and B. Sudret. “A unified framework for multilevel uncertainty quantification in Bayesian inverse problems”. In: *Probabilistic Engineering Mechanics* 43 (2016), pp. 68–84. DOI: 10.1016/j.probengmech.2015.09.007.

[3] J. B. Nagel and B. Sudret. “Hamiltonian Monte Carlo and Borrowing Strength in Hierarchical Inverse Problems”. In: *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 2.3, B4015008 (2016), pp. 1–12. DOI: 10.1061/AJRUA6.0000847.

[4] J. B. Nagel and B. Sudret. “Spectral likelihood expansions for Bayesian inference”. In: *Journal of Computational Physics* 309 (2016), pp. 267–294. DOI: 10.1016/j.jcp.2015.12.047.

1.2.2 Conference proceedings

[5] J. B. Nagel and B. Sudret. “Probabilistic Inversion for Estimating the Variability of Material Properties: A Bayesian Multilevel Approach”. In: *11th International Probabilistic Workshop (IPW11)*. Ed. by D. Novák and M. Vořechovský. Brno, Czech Republic: Litera, 2013, pp. 293–303. DOI: 10.3929/ethz-a-010034843.

[6] J. B. Nagel and B. Sudret. “A Bayesian Multilevel Framework for Uncertainty Characterization and the NASA Langley Multidisciplinary UQ Challenge”. In: *16th AIAA Non-Deterministic Approaches Conference (SciTech 2014)*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2014. DOI: 10.2514/6.2014-1502.

[7] J. B. Nagel and B. Sudret. “A Bayesian Multilevel Approach to Optimally Estimate Material Properties”. In: *2nd International Conference on Vulnerability and Risk Analysis and Management and 6th International Symposium on Uncertainty Modeling and Analysis (ICVRAM & ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 151, pp. 1504–1513. DOI: 10.1061/9780784413609.151.

[8] J. B. Nagel, N. Mojsilovic, and B. Sudret. “Bayesian Assessment of the Compressive Strength of Structural Masonry”. In: *12th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP12)*. Vancouver, Canada: University of British Columbia, 2015. DOI: 10.14288/1.0076072.

1.2.3 Other presentations

[9] J. B. Nagel and B. Sudret. "Bayesian Multilevel Model Calibration for Inversion of "Perfect" Data in the Presence of Uncertainty". In: *MascotNum Workshop on Computer Experiments and Meta-models for Uncertainty Quantification (MascotNum 2014)*. Zürich, Switzerland, April 2014.

[10] J. B. Nagel and B. Sudret. "PCE-Metamodeling for Inverse Heat Conduction". In: *1st Pan-American Congress on Computational Mechanics (PANACM 2015)*. Buenos Aires, Argentina, April 2015.

[11] J. B. Nagel and B. Sudret. "Optimal Transportation for Bayesian Inference in Engineering". In: *International Symposium on Reliability of Engineering Systems (SRES 2015)*. Hangzhou, China, October 2015.

[12] J. B. Nagel and B. Sudret. "Spectral Likelihood Expansions and Nonparametric Posterior Surrogates". In: *SIAM Conference on Uncertainty Quantification (SIAM UQ 2016)*. Lausanne, Switzerland, April 2016.

[13] J. B. Nagel and B. Sudret. "Nonparametric posterior surrogates based on spectral likelihood expansions and least angle regression". In: *European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2016)*. Crete Island, Greece, June 2016.

1.3 Outline

An overview of the doctoral thesis and its structure is now given. The document is basically divided into three parts. It starts with basic introductions to uncertainty quantification and Bayesian inference in Part I. The key contributions of the research work in form of the most important publications are compiled in Part II. Some further unpublished investigations as well as a detailed hydrological case study are conducted in Part III. A short overview of how the main topics and the associated publications are organized in parts and chapters is tabulated below. Detailed chapter summaries follow directly thereafter.

Part I	Chapter 2	Uncertainty quantification	
	Chapter 3	Bayesian inference	
Part II	Chapter 4	Aleatory variability	[2]
	Chapter 5	Hamiltonian Monte Carlo	[3]
	Chapter 6	NASA Langley challenge	[1]
	Chapter 7	Structural masonry	[8]
	Chapter 8	Spectral Bayesian inference	[4]
Part III	Chapter 9	Optimal transportation	
	Chapter 10	Hydrological model calibration	

1.3.1 Elementary introductions

The thesis starts with two introductory chapters on uncertainty quantification and Bayesian inference in engineering problems. This material provides the necessary background information as well as complementary perspectives on the more advanced developments that follow. Forward and inverse problems are discussed within the framework of probabilistic uncertainty quantification. State-of-the-art techniques for the computational forward and backward propagation of uncertainty are reviewed.

An introduction to uncertainty quantification with a clear focus on probabilistic methods and forward propagation is provided in Chapter 2. The quantitative characterization of the response distribution of a mechanical model due to randomness in the input parameters, e.g. material properties, object geometry, environmental loads or operating conditions, is the classical example problem. Monte Carlo simulation, Taylor series expansions and more global metamodeling techniques are presented in this context. The latter includes stochastic spectral methods such as polynomial chaos expansions which are used throughout the whole dissertation. Non-intrusive computations based on a linear least squares minimization problem and its ordinary least squares solution are concentrated on. The curse of dimensionality as well as the hope for sparsity are discussed.

Chapter 3 contains an elementary introduction to the Bayesian data analysis of engineering systems. This offers a principled way of quantifying and reducing epistemic parameter uncertainties. Experimental data that are only indirectly associated to the actual quantities of interest are analyzed to that end. An example is the determination of actually uncertain properties of a material with measurements of its behavior under certain test conditions. Basic inferential principles founded on the likelihood function as well as the prior and the

posterior distribution are introduced. More advanced topics such as evidence-based model comparison and some practical issues related to the parametrization of statistical models are also covered. Conventional approaches to computational Bayesian inference based on random sampling or mathematical optimization are surveyed, e.g. Markov chain Monte Carlo and variational inference. The convenient calculation of the extremely large or small quantities that typically arise in Bayesian computations is discussed. Bayesian inverse problems are dealt with in greater detail together with related issues such as the quantification of model prediction error.

1.3.2 Aleatory variability

While the Bayesian solution to inverse problems satisfactorily accounts for epistemic types of uncertainty, it does not allow for the incorporation of aleatory types. Two examples of this form of uncertainty are the stochastic variation of the environmental or operating conditions over time and the randomness within an ensemble of structural elements due to manufacturing tolerances. This is a major limitation and motivates the research question of how to deal with aleatory input variability in Bayesian inverse problems. The answer to the question is a core topic in the dissertation and occupies two chapters at the very least.

A hierarchical framework for managing heterogeneous types of uncertainty in Bayesian inverse problems is proposed in Chapter 4. The formulation rests on multilevel models that interrelate different system components through deterministic simulators and conditional probability distributions. It allows one to reduce the epistemic uncertainty of unknowns that are fixed yet unknown and to identify the distribution of quantities that vary throughout a series of experiments. Random measurement noise and aleatory nuisance variables are taken into account at the same time. All available sources of information such as experimental data and expert knowledge can be harnessed and optimally combined. This is especially important in civil engineering applications where information is scarce and uncertainty dominates. The framework is demonstrated and its computational challenges are identified through estimating the material variability across equally manufactured structural elements. Inference can be either based on a low-dimensional formulation with an integrated likelihood function or on a high-dimensional variant with many unknowns.

After the framework for Bayesian inversion under epistemic and aleatory uncertainty has been elaborated, specialized solvers have to be implemented for computing the posterior efficiently. In Chapter 5 we propose Hamiltonian Monte Carlo in order to cope with the practical difficulties of high-dimensional Bayesian multilevel modeling. As a member from the extended Markov chain Monte Carlo family, the algorithm explores the posterior in a sampling-based manner. The idea is to embed the space of the unknown parameters in an auxiliary space and to perform the Markovian updates in such a way that they mix well in the original space of interest. This principle is inspired by systems and concepts from classical and statistical mechanics, i.e. Hamiltonian dynamics and the Boltzmann distribution. It calls for derivatives of the log–posterior density and the forward model. Hamiltonian Monte Carlo is shown to be a highly efficient solver for inverse problems under uncertainty and variability. It drastically outperforms a random walk Metropolis algorithm in a benchmark problem with more than hundred unknowns. The posterior can be sampled almost independently.

1.3.3 Computational methods

Beyond the treatment of aleatory variability in inverse problems, the development of novel methods for computational Bayesian inference is another central theme of the thesis. Although Hamiltonian Monte Carlo is an attractive algorithm, it does not overcome the principal limitations of sampling techniques in general. Fundamental alternatives to Markov chain and sequential Monte Carlo are thus needed. Two entirely different approaches to compute the posterior distribution numerically are developed and investigated. They are based on approximations of the posterior probability density function or correspondingly distributed random variables. The identification of the thermal properties of a composite material with inclusions poses a comparably simple inverse heat conduction problem that serves testing and demonstration purposes.

Spectral Bayesian inference is developed in Chapter 8 as a completely new and pretty elegant technique for posterior computations. The main idea is to decompose the likelihood function into a converging series of orthogonal polynomials. If orthogonality is defined with respect to the prior weight, this spectral likelihood expansion has some surprisingly interesting properties. It gives rise to a nonparametric representation of the normalized posterior density and enables semi-analytic and sampling-free inference. The model evidence as well as the posterior moments are related to the expansion coefficients from which they can be easily extracted. Posterior uncertainty propagation through general computational models can be accomplished based on prior polynomial chaos expansions. It is proposed to compute the expansion coefficients by a discrete linear least squares projection. A perturbation-theoretic interpretation of the orthogonal series expansion of the posterior suggests a change of the reference density from the prior to an auxiliary weight function. This improves the accuracy and efficiency of the spectral method dramatically. The advantages and shortcomings of spectral

Bayesian inference are highlighted by reference to classical distribution fitting and the inverse heat conduction problem.

Another approach to computational Bayesian inference that was recently devised by Prof. Youssef Marzouk and his group at MIT is investigated in Chapter 9. It is based on the translocation of probability mass from the prior to the posterior measure. A function of random variables distributed according to the prior is constructed in such a way that the transformed variables follow the posterior. One can establish a connection between variational Bayesian inference and this transport-based formulation. This permits to compute the posterior by solving an optimization problem with an information-theoretic optimality criterion. The random variable transformation is parametrized through multivariate polynomials up to a certain degree. After the computation of a suitable transform, one can draw independent and equally weighted samples from the posterior. This compelling feature distinguishes the approach from conventional sampling techniques.

1.3.4 Practical applications

A number of inverse problems involving real data and forward models are solved with the previously developed methods towards the end of the thesis. This can be seen as a justification and appreciation of the more formal developments, but should not hide the fact that it actually motivated some of them in the first place. Spectral Bayesian inference for instance originated in the context of the NASA Langley multidisciplinary uncertainty quantification challenge. The initial intention to use a polynomial approximation of the log–likelihood function in conjunction with Markov chain Monte Carlo sampling has evolved into the idea for a spectral likelihood expansion which renders further posterior sampling completely unnecessary.

Chapter 6 is the outcome of participating in the NASA Langley uncertainty quantification challenge in 2013–2014. The challenge contained a set of interlinked uncertainty quantification problems from the domain of aerospace engineering. In this chapter the calibration sub-problem is interpreted and solved in the developed framework of Bayesian multilevel modeling. A black-box model describing the behavior of a miniature civilian aircraft under adverse flight conditions and associated data were provided by NASA. The primary goal was the reduction of epistemic uncertainty of the model parameters that are fixed yet unknown and the identification of the hyperparameters that determine the distribution of the aleatory variables. Due to some peculiarities of the problem statement related to a zero-noise or perfect data condition, the likelihood function only arises as the solution to a secondary uncertainty forward propagation problem. For that reason it cannot be evaluated exactly. A statistical approximation of the likelihood based on Monte Carlo simulation and kernel density estimation is therefore proposed. Employing this biased and noisy likelihood estimator for sampling the posterior via Markov chain Monte Carlo alters the Metropolis–Hastings transition kernel. The induced modifications on the posterior level are investigated and mitigated by means of partial data augmentation.

Bayesian multilevel modeling also facilitates problem-solving in structural masonry. A hierarchical approach to assess the compressive strength of masonry walls is presented in Chapter 7. Many current methods suffer from their homogeneous treatment of the composite material or simply fail in uncertainty quantification. Other standardized methods overpredict the compressive strength to an alarming extent. The devised approach allows one to improve the accuracy of the predictions and assess their quality. It models structural masonry heterogeneously and quantifies the arising uncertainties consistently. System-level data related to the masonry wall specimens and component-level data of the brick units and mortar are analyzed jointly. The experimental data were collected in a series of compressive tests performed by Dr. Nebojsa Mojsilovic and his students in the laboratories of the Institute of Structural Engineering (IBK) at the ETH Zürich. After the calibration of the unknown parameters and hyperparameters, one can probabilistically predict the compressive strength based on measurements of the constituent ensembles used.

Another real-world problem is solved in Chapter 10 where a hydrological urban drainage simulator is calibrated. Epistemic parameter uncertainties are reduced while random measurement noise and systematic modeling errors are anticipated and statistically identified. This allows for a thorough treatment of the emerging sources of error and uncertainty in the dynamical simulation of water systems. It also suggests the possibility for model correction. The catchment area of Adliswil, a municipality located around the river Sihl at the southern end of the city Zürich, is studied during a rainfall event. Experimental data and training runs of the simulator were provided by the Swiss Federal Institute of Aquatic Science and Technology (Eawag) in Dübendorf. Advanced techniques for dimension reduction, surrogate modeling and stochastic sampling are combined to this effect. Principal component analysis allows us to reduce the output dimensionality of the deterministic simulator that predicts a whole times series. Sparse polynomial chaos expansions are subsequently used in order to emulate the input-output relationship the forward model defines. The posterior distributions of two different Bayesian models are sampled via Markov chain Monte Carlo techniques and compared with each other.

Part I

Introduction

Chapter 2

Uncertainty quantification

Modern engineering and scientific computing stimulate each other in a fruitful way. This synergy has led to a number of great discoveries and breakthroughs, e.g. the *Monte Carlo method* [1] and the *Kalman filter* [2, 3]. Due to the steady increase and availability of computer capacities on the one hand and algorithmic advances on the other hand, progress in computational science and engineering is nowadays made at an unprecedented pace. Mature programming environments and ready-made software packages are available for a variety of dedicated tasks, e.g. for finite element analysis and multiphysics simulation. The necessary processing power and computing time are provided by personal computers or high-performance computing clusters. More and more complex systems can be simulated in an ever-increasing level of detail. This trend continues to the present day.

The accurate simulation of large-scale systems typically hinges on the knowledge of a great number of physical parameters. In practice they are hardly known exactly, though. Even if all parameters of a certain model could be specified somehow, this would not necessarily shield from systematic errors and guarantee satisfactory predictions. Inadequacies are immanent in all physical models to some degree, e.g. due to missing physics or unresolved scales. The investigation of different sources and levels of inaccuracy and imprecision in numerical simulations is therefore suggested. This is the objective of *uncertainty quantification* (UQ) [4, 5].

In a wider sense, UQ deals with all uncertainties of computer simulations within an academic or industrial context. One certainly encounters various quite different sources of error in scientific computing [6, 7]. These include but are not limited to *parameter uncertainty* [8, 9], *numerical inaccuracy* [10–12] and *measurement noise* [13–15]. While the two latter types of errors are treated in statistical and numerical analysis, UQ concentrates on the first-mentioned type in a narrower sense. With this ambition, UQ has recently emerged as an active research field at the intersection of statistics, applied mathematics, computer science and engineering.

In engineering applications, one commonly distinguishes between *epistemic uncertainty* and *aleatory variability* [16]. The former refers to a lack of knowledge of the analyst, whereas the latter relates to a natural randomness of the system. In statistical inference, *frequentist* and *Bayesian* interpretations of probability differ in the way they address these uncertainties [17–19]. On the one hand, probabilities are only employed for describing objective frequencies. On the other hand, they are also utilized for representing subjective ignorance. While it is enjoyable to reflect and dispute about such a categorization, with good reason one may wonder whether it is really helpful or rather creates confusion. Taking a pragmatic point of view and declining the related philosophical debates, the use of probability theory for either uncertainty is a common modeling choice in UQ.

While the fundamentals of UQ are therefore well-established in principle, the actual challenge is the complexity of modern engineering problems. The system under study is often composed of different interacting components, each of which may be already complex taken only by itself. A complete model of this system usually consists of numerous deterministic and stochastic sub-models of the physics and uncertainty involved. Abstractly one may speak of the "model universe" [20] or "theoretical world" [21] that embodies all assumptions and idealizations of that overall model. In Fig. 2.1 it is attempted to illustrate this conception. The goal of UQ then becomes the quantitative analysis and global management of uncertainty throughout the integrated system.

In a real-case scenario, the workflow typically involves a whole chain of intertwined UQ analyses. Broadly speaking, these tasks can be divided into *forward UQ*, i.e. characterizing the model outputs, and *inverse UQ*, i.e. learning about the model inputs. See Fig. 2.2 for a visualization. In *uncertainty propagation* [22, 23] one tries to find the full distribution of model outputs for given input uncertainties. This includes *reliability analysis* [24, 25] where one focuses on the computation of the failure probability, e.g. that the system output exceeds a certain threshold. While these are forward UQ problems, *parameter identification* [26, 27] and *data assimilation* [28, 29] belong to inverse UQ. Given noisy observations of the system output, the goal is the experimental estimation of unknown system parameters or dynamical states.

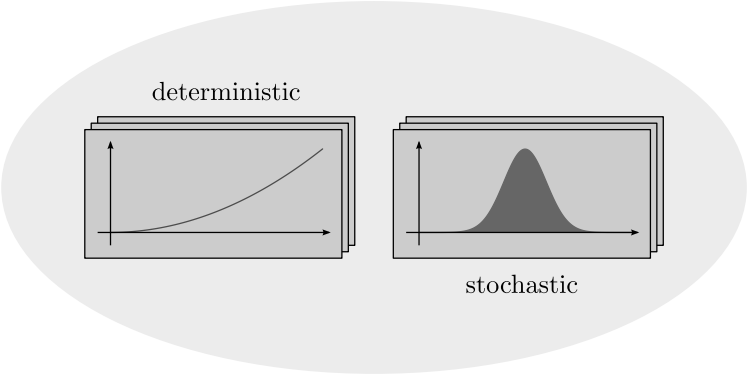


Figure 2.1: Model universe.

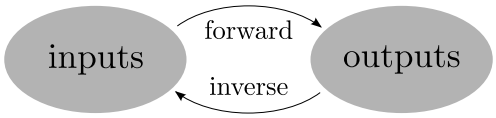


Figure 2.2: Forward and inverse problems.

There are also some intermediate UQ tasks that concentrate on understanding the input-output relationship that the model establishes. By mimicking this relation or exploiting its structure one can accelerate forward as well as inverse UQ problems. *Surrogate modeling* [30, 31] aims at an approximation of the model that is both easy to interpret and cheap to evaluate. Similarly, *model reduction* [32, 33] subsumes techniques that try to simplify the model while a reasonable degree of accuracy is maintained. In *sensitivity analysis* [34, 35] one compares and ranks the importance of different inputs with respect to their effect on the outputs. This allows one to identify the most and least influential parameters. While many of the activities in between forward and inverse UQ are worthwhile by themselves, they still need to be seen in the bigger picture of *risk analysis* [36, 37] and *decision making* [38, 39].

The remainder of this introductory chapter is organized as follows. Some fundamentals of uncertainty propagation are reviewed in Section 2.1. A discussion about local methods such as Taylor series approximations follows in Section 2.2. This eventually motivates global surrogate modeling approaches based on orthogonal polynomials in Section 2.3. Least squares minimization techniques for computing function approximations are subsequently presented in Section 2.4. Multivariate model outputs are considered in Section 2.5. Issues related to high-dimensionality and sparsity are finally discussed in Section 2.6.

2.1 Uncertainty propagation

We now focus on probabilistic uncertainty propagation. The goal is to quantify the influence of input parameter uncertainty on the predictions of an engineering model [40, 41]. By representing the uncertain inputs as random variables with a prespecified probability distribution, the problem becomes to characterize the corresponding output distribution. After a short intermezzo with slightly more technical probability theory, only a basic familiarity with elementary statistics is required. Introductions for an engineering target audience can be found in [42–44].

2.1.1 Engineering model

In the context of UQ, a *model* of an engineering system is a mathematical representation or computational simulation of the relevant physical processes. For given input parameters $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ from the domain $\mathcal{D}_{\boldsymbol{x}} \subseteq \mathbb{R}^M$ with $M \in \mathbb{N}_{>0}$, the model predicts an output of interest $\tilde{y} \in \mathbb{R}$. A single response quantity is considered here for the sake of simplicity. The extension to multivariate outputs is straightforward, though. Accordingly, the model can be thought of as a scalar-valued function

$$\begin{aligned} \mathcal{M}\colon \mathcal{D}_{\boldsymbol{x}} &\to \mathbb{R} \\ \boldsymbol{x} &\mapsto \tilde{y} = \mathcal{M}(\boldsymbol{x}). \end{aligned} \tag{2.1}$$

Many different types of such predictive models are encountered in engineering problems. This includes simple analytic expressions as well as numerical solutions of the governing equations. Especially in the latter case, the model symbolized in Eq. (2.1) is often treated as a *black-box*, i.e. it is only evaluated in a pointwise manner. Its internal structure may be not known, too complex or simply not considered explicitly. The only requirement is that the model is available in an executable form.

2.1.2 Input distribution

Provided that the simulator captures the main characteristics of the system well in general, one can calculate the output $\tilde{y} = \mathcal{M}(\boldsymbol{x})$ for arbitrary input values $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$. This allows for an accurate forecast in case the input parameters best describing a scenario are exactly known. In other circumstances the inputs are uncertain, e.g. owing to a lack of knowledge or a natural variability. Then one can represent them as a random vector

$$\boldsymbol{X} \sim \pi(\boldsymbol{x}). \tag{2.2}$$

For the time being, we assume that the density function $\pi(\boldsymbol{x})$ of the input distribution in Eq. (2.2) is already given. It is remarked that the systematic specification of this density with output measurements is indeed the core subject in this thesis. An introduction to the reduction of epistemic uncertainty is provided in Chapter 3. The quantification of aleatory variability is the thematic priority of Chapters 4 and 5.

The input distribution is often characterized through its first statistical moments, e.g. its mean vector $\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}]$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathrm{Cov}[\boldsymbol{X}]$. These moments are assumed to be well-defined and finite throughout the dissertation. They are given as

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}] = \int_{\mathcal{D}_{\boldsymbol{x}}} \boldsymbol{x}\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{2.3}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})\,(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^{\top}\right] = \int_{\mathcal{D}_{\boldsymbol{x}}} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})\,(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})^{\top}\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{2.4}$$

The mean and covariance in Eqs. (2.3) and (2.4) are often taken as measures of the location and dispersion of the input distribution, e.g. they indicate the typical value and the variation of the uncertain inputs.

2.1.3 Probability theory

Rigorous probability theory is often perceived as somewhat counterintuitive by practitioners of calculus-based probability. Random variables are actually functions, integration is introduced before differentiation and the conditional expectation is a prerequisite for the conditional distribution. In order to clarify some of the fundamental notions that are behind Eq. (2.2) and that are used throughout the whole thesis, we dare a brief review of probability spaces, random variables, expectation values and density functions.

Let us consider a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. The triplet consists of a sample space Ω of random outcomes, a σ-field $\mathcal{F} \subseteq 2^{\Omega}$ and a probability measure $\mathbb{P}\colon \mathcal{F} \to [0,1]$. A *random vector* on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathcal{D}_{\boldsymbol{x}}$ is a measurable function $\boldsymbol{X}\colon (\Omega, \mathcal{F}) \to (\mathcal{D}_{\boldsymbol{x}}, \mathcal{B}(\mathcal{D}_{\boldsymbol{x}}))$. This means that $\boldsymbol{X}^{-1}(B) = \{\omega \in \Omega | \boldsymbol{X}(\omega) \in B\} \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathcal{D}_{\boldsymbol{x}})$ in the Borel σ-algebra $\mathcal{B}(\mathcal{D}_{\boldsymbol{x}})$ on $\mathcal{D}_{\boldsymbol{x}}$. The random vector induces a so-called *image law* or *probability distribution* on $(\mathcal{D}_{\boldsymbol{x}}, \mathcal{B}(\mathcal{D}_{\boldsymbol{x}}))$ by

$$\mathbb{P}_{\boldsymbol{X}}(B) = \mathbb{P} \circ \boldsymbol{X}^{-1}(B). \tag{2.5}$$

One can write $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{\boldsymbol{X}} (\mathcal{D}_{\boldsymbol{x}}, \mathcal{B}(\mathcal{D}_{\boldsymbol{x}}), \mathbb{P}_{\boldsymbol{X}})$ in order to summarize this basic probability setup. Note that only spaces and mappings have been introduced as yet.

A *quantity of interest* (QoI) is a scalar-valued Borel function $h\colon (\mathcal{D}_{\boldsymbol{x}}, \mathcal{B}(\mathcal{D}_{\boldsymbol{x}})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. It defines a $\mathbb{R}$-valued random variable $h(\boldsymbol{X}) = h \circ \boldsymbol{X}$ with a distribution $\mathbb{P}_h = \mathbb{P}_{\boldsymbol{X}} \circ h^{-1}$. The *expectation value* of this random variable is defined as the Lebesgue integral

$$\mathbb{E}[h(\boldsymbol{X})] = \int_{\Omega} h(\boldsymbol{X}(\omega))\, \mathbb{P}(\mathrm{d}\omega) = \int_{\mathcal{D}_{\boldsymbol{x}}} h(\boldsymbol{x})\, \mathbb{P}_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}) = \int_{\mathbb{R}} h'\, \mathbb{P}_h(\mathrm{d}h'). \tag{2.6}$$

Similarly, the integration of vector-valued functions is treated in componentwise manner. It is remarked that the probabilities in Eq. (2.5) can be expressed as expectation values of the form as in Eq. (2.6) by $\mathbb{P}_{\boldsymbol{X}}(B) = \mathbb{E}[\mathrm{I}_B(\boldsymbol{X})] = \int_B \mathbb{P}_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x})$. Here, $\mathrm{I}_B\colon \Omega \to \{0,1\}$ is the indicator function of the set B with $\mathrm{I}_B(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in B$ and $\mathrm{I}_B(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \notin B$.

A *probability density function* (PDF) of $\mathbb{P}_{\boldsymbol{X}}$ with respect to the Lebesgue measure is any measurable function $\pi\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R}^+$ for which the expectation $\mathbb{E}[h(\boldsymbol{X})]$ can be written as

$$\mathbb{E}[h(\boldsymbol{X})] = \int_{\mathcal{D}_{\boldsymbol{x}}} h(\boldsymbol{x})\, \mathbb{P}_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}) = \int_{\mathcal{D}_{\boldsymbol{x}}} h(\boldsymbol{x})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{2.7}$$

Similar to Eq. (2.7), also the probabilities $\mathbb{P}_{\boldsymbol{X}}(B)$ can be rewritten in a way involving the density function as $\mathbb{P}_{\boldsymbol{X}}(B) = \int_B \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$. Note that PDFs are only implicitly defined via these integral relations.

Assuming that the random variable $\boldsymbol{X}$ has the image law $\mathbb{P}_{\boldsymbol{X}}$ with a PDF $\pi(\boldsymbol{x})$, the relevant expectation values in Eq. (2.6) can be determined as per Eq. (2.7). That all is behind Eq. (2.2). In sum, a PDF is a convenient way to specify a whole probability distribution. Hence, we use this PDF-oriented language and notation throughout the thesis.

2.1.4 Output distribution

In uncertainty propagation one tries to quantify the distribution of the model outputs that results from the randomness in the inputs. The simulator is therefore treated as a measurable function or QoI $\mathcal{M}\colon (\mathcal{D}_{\boldsymbol{x}}, \mathcal{B}(\mathcal{D}_{\boldsymbol{x}})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. One considers the response random variable defined by

$$\tilde{Y} = \mathcal{M}(\boldsymbol{X}). \tag{2.8}$$

This setup is summarized as $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{\boldsymbol{X}} (\mathcal{D}_{\boldsymbol{x}}, \mathcal{B}(\mathcal{D}_{\boldsymbol{x}}), \mathbb{P}_{\boldsymbol{X}}) \xrightarrow{\mathcal{M}} (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_{\tilde{Y}})$. An illustration of uncertainty forward propagation is provided in Fig. 2.3. The input distribution $\mathbb{P}_{\boldsymbol{X}}$ and the push-forward measure $\mathbb{P}_{\tilde{Y}} = \mathbb{P}_{\boldsymbol{X}} \circ \mathcal{M}^{-1}$ therein are characterized by their probability densities. The response distribution can be complex, which is exemplified by two different modes. A short look at Fig. 3.2 allows one to catch a glimpse at uncertainty backpropagation already.

Similar as for the moments of the inputs in Eqs. (2.3) and (2.4), the output distribution is often simply summarized by the mean $\mu_{\tilde{Y}} = \mathbb{E}[\mathcal{M}(\boldsymbol{X})]$ and the variance $\sigma^2_{\tilde{Y}} = \mathrm{Var}[\mathcal{M}(\boldsymbol{X})]$ of the random variable in Eq. (2.8). Herein, their existence and finiteness is always presumed. These moments are respectively given as

$$\mu_{\tilde{Y}} = \mathbb{E}[\mathcal{M}(\boldsymbol{X})] = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{M}(\boldsymbol{x})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{2.9}$$

$$\sigma^2_{\tilde{Y}} = \mathbb{E}\left[\left(\mathcal{M}(\boldsymbol{X}) - \mu_{\tilde{Y}}\right)^2\right] = \int_{\mathcal{D}_{\boldsymbol{x}}} \left(\mathcal{M}(\boldsymbol{x}) - \mu_{\tilde{Y}}\right)^2 \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{2.10}$$

For simple problems where the random response is unimodal and not too far from being Gaussian, the mean and variance in Eqs. (2.9) and (2.10) can be taken as measures of the location and scale. Summarizing the shape of more complex distributions may very well require higher moments such as skewness and kurtosis, though. When the output distribution is multimodal such as in Fig. 2.3, the first statistical moments are difficult to interpret. In this case the distribution can be meaningfully characterized through its full density only.

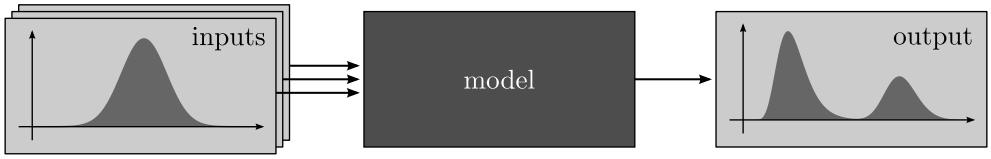


Figure 2.3: Forward uncertainty propagation.

2.1.5 Monte Carlo simulation

An appealingly simple approach to compute the first moments of the output distribution is Monte Carlo (MC) simulation. For $K \in \mathbb{N}_{>1}$ representative input samples $\mathcal{X} = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(K)})$, which are independently drawn from the input distribution in Eq. (2.2), one has to compute the corresponding responses $\mathcal{Y} = (\tilde{y}^{(1)}, \dots, \tilde{y}^{(K)})^\top$. Here, $\tilde{y}^{(k)} = \mathcal{M}(\boldsymbol{x}^{(k)})$ are realizations of the random variable in Eq. (2.8) for $k = 1, \dots, K$. The moments in Eqs. (2.9) and (2.10) can then be estimated via the sample approximations

$$\overline{\mu}_{\tilde{Y}} = \frac{1}{K} \sum_{k=1}^{K} \tilde{y}^{(k)}, \quad \overline{\sigma}^2_{\tilde{Y}} = \frac{1}{K-1} \sum_{k=1}^{K} \left(\tilde{y}^{(k)} - \overline{\mu}_{\tilde{Y}}\right)^2. \tag{2.11}$$

This is a universal and solid approach to characterize the response distribution. Because the MC estimates in Eq. (2.11), provided that at least the first two and four moments respectively exist, enjoy input dimension–independent convergence rates of the statistical sampling errors with the number of random samples, they are often used in high-dimensional problems. Especially for problems of low and moderate dimension, other alternatives might be superior. A local method based on a Taylor expansion of the model and more global metamodeling techniques are discussed in Sections 2.2 and 2.3, respectively.

2.2 Taylor approximation

In the error analysis of physical experiments, one should always quantify the influence of measurement uncertainties on the final results [13–15]. In this context one is often interested in a simple function of the observed data. A commonly encountered method of *error propagation* is then based on a low-order Taylor approximation of this function. This approach is certainly appealing in simple cases. In principle, one may also contemplate it for the uncertainty propagation through general engineering models. That is at the basis of the *stochastic perturbation method* [45]. A few introductory remarks are provided hereafter.

Given that the model $\mathcal{M}(\boldsymbol{x})$ is sufficiently differentiable, one can approximate it through a truncated Taylor series around a value $\boldsymbol{x}_0 \in \mathcal{D}_{\boldsymbol{x}}$ from its domain. Since the approximation will be only accurate in some neighborhood of the expansion point, one chooses the mean value $\boldsymbol{\mu}_{\boldsymbol{X}} = (\mu_{X_1}, \dots, \mu_{X_M})^\top$ in Eq. (2.3). The second-order Taylor approximation of $\mathcal{M}(\boldsymbol{x})$ about this point $\boldsymbol{x}_0 = \boldsymbol{\mu}_{\boldsymbol{X}}$ is then given as

$$\mathcal{M}(\boldsymbol{x}) \approx \mathcal{M}(\boldsymbol{\mu}_{\boldsymbol{X}}) + \sum_{i=1}^{M} \frac{\partial \mathcal{M}}{\partial x_i}\bigg|_{\boldsymbol{\mu}_{\boldsymbol{X}}} (x_i - \mu_{X_i}) + \frac{1}{2} \sum_{i,j=1}^{M} \frac{\partial^2 \mathcal{M}}{\partial x_i \partial x_j}\bigg|_{\boldsymbol{\mu}_{\boldsymbol{X}}} (x_i - \mu_{X_i})(x_j - \mu_{X_j}). \tag{2.12}$$

With this, one can find the corresponding second-order approximation of the expected value $\mu_{\tilde{Y}} = \mathbb{E}[\mathcal{M}(\boldsymbol{X})]$ of the model output in Eq. (2.9). A simple calculation yields

$$\mu_{\tilde{Y}} \approx \mathcal{M}(\boldsymbol{\mu}_{\boldsymbol{X}}) + \frac{1}{2} \sum_{i,j=1}^{M} \frac{\partial^2 \mathcal{M}}{\partial x_i \partial x_j}\bigg|_{\boldsymbol{\mu}_{\boldsymbol{X}}} \mathrm{Cov}[X_i, X_j]. \tag{2.13}$$

For independent inputs with $\mathrm{Cov}[X_i, X_j] = 0$ whenever $i \neq j$, the approximation simply becomes $\mu_{\tilde{Y}} \approx \mathcal{M}(\boldsymbol{\mu}_{\boldsymbol{X}}) + \frac{1}{2} \sum_{i=1}^{M} \frac{\partial^2 \mathcal{M}}{\partial x_i^2}|_{\boldsymbol{\mu}_{\boldsymbol{X}}} \mathrm{Var}[X_i]$. Similarly, the first-order approximation of the model output variance $\sigma^2_{\tilde{Y}} = \mathrm{Var}[\mathcal{M}(\boldsymbol{X})]$ in Eq. (2.10) can be written as

$$\sigma^2_{\tilde{Y}} \approx \sum_{i,j=1}^{M} \frac{\partial \mathcal{M}}{\partial x_i}\bigg|_{\boldsymbol{\mu}_{\boldsymbol{X}}} \frac{\partial \mathcal{M}}{\partial x_j}\bigg|_{\boldsymbol{\mu}_{\boldsymbol{X}}} \mathrm{Cov}[X_i, X_j]. \tag{2.14}$$

The well-known law of error propagation $\sigma^2_{\tilde{Y}} \approx \sum_{i=1}^{M} (\frac{\partial \mathcal{M}}{\partial x_i}|_{\boldsymbol{\mu}_{\boldsymbol{X}}})^2 \mathrm{Var}[X_i]$ is a consequence of independent inputs. Of course, one could calculate Taylor approximations of higher order than in Eqs. (2.12) to (2.14).

On the downside, these formulas are only accurate in very simple situations, e.g. for unimodal input distributions for which the expected value indeed indicates most of the probability mass and for models that are very smooth over the typical input variation. The applicability of the method is also limited due to the necessity to compute partial derivatives. All things considered, local Taylor series approximations do not establish a reliable and viable alternative to more global attempts.

2.3 Surrogate modeling

Another approach is to construct a global *surrogate* or *metamodel* of the simulator in the vein of a *response surface* [46, 47]. The metamodel is built so as to emulate or mimic the behavior of the original model over its whole domain. Typically it is cheap to evaluate and can therefore supersede the full model in analyses that require many model runs, e.g. uncertainty propagation, sensitivity analysis and parameter estimation. This renders the analysis of systems possible where using the simulator is ruled out due to the incurred computational cost. Of course, this only works subject to the condition that the cost of computing a sufficiently accurate surrogate does not exceed the available budget either.

Widespread classes of metamodels are based on polynomial chaos expansions [48, 49], Gaussian process models [50, 51], artificial neural networks [52, 53] and support vector machines [54, 55]. Traditionally these emulator types have been developed and publicized in different scientific communities and disciplines. While the two first-mentioned techniques are mainly used in engineering and statistics, respectively, the two last-mentioned ones are rooted in machine learning. Nowadays they are used in a more cross-disciplinary and problem-oriented manner. Polynomial chaoses [56, 57], Gaussian processes [58, 59] and neural networks [60, 61] are used in numerous different ways for parameter estimation. Each approach certainly has its own advantages and the performance depends on the specific problem at hand. One can combine different techniques in order to improve the efficiency [62–65].

In the following we concentrate on polynomial chaos–based metamodels. Even though the theoretical foundations date back to the first half of the last century [66, 67], this type of stochastic spectral expansion

was popularized only during the last decades in the context of stochastic finite elements [68, 69]. Technically speaking, a random variable is expanded with respect to a number of basis random variables. We approach the topic from the perspective of function approximation [70, 71]. This matches more closely how the surrogate model is used after its construction in practice, i.e. it approximates function values for more or less arbitrarily chosen inputs.

We start with a scalar-valued model $\mathcal{M}\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R}$ as in Eq. (2.1). It maps inputs $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ to outputs $\tilde{y} = \mathcal{M}(\boldsymbol{x}) \in \mathbb{R}$. In the vector-valued case one would treat each component separately. Alternatively, one could make use of a non-canonical basis representation of the model output space and subsequently consider the coefficients individually. This is further elaborated on in Section 2.5.

2.3.1 Spectral expansions

Let us consider attractive spaces for the simulator. Given a weight function $w\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R}^+$, one may try to justify or simply assume that the model $\mathcal{M} \in L^2_w(\mathcal{D}_{\boldsymbol{x}})$ belongs to the function space

$$L^2_w(\mathcal{D}_{\boldsymbol{x}}) = \left\{ u\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R} \middle| \int_{\mathcal{D}_{\boldsymbol{x}}} u^2(\boldsymbol{x})\, w(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} < \infty \right\}. \tag{2.15}$$

This is a Hilbert space with an inner product $\langle \cdot, \cdot \rangle_w$ and an associated norm $\|\cdot\|_w$. For any two elements $u, v \in L^2_w(\mathcal{D}_{\boldsymbol{x}})$ these are defined as

$$\langle u, v \rangle_w = \int_{\mathcal{D}_{\boldsymbol{x}}} u(\boldsymbol{x}) v(\boldsymbol{x})\, w(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \quad \|u\|_w = \langle u, u \rangle_w^{1/2}. \tag{2.16}$$

It now seems natural to seek for a Hilbert basis with respect to which one can expand vectors. Let $\{\Psi_i\}_{i \in \mathbb{N}_{>0}}$ be a complete orthonormal set of vector space elements $\Psi_i \in L^2_w(\mathcal{D}_{\boldsymbol{x}})$. Thus for all $i, j \in \mathbb{N}_{>0}$ one has $\langle \Psi_i, \Psi_j \rangle_w = \delta_{ij}$, where δ_{ij} is the Kronecker delta. As an element of the space in Eq. (2.15) the model can be expanded with respect to the orthonormal basis $\{\Psi_i\}_{i \in \mathbb{N}_{>0}}$. This so-called *spectral expansion* is given as

$$\mathcal{M} = \sum_{i=1}^{\infty} a_i \Psi_i, \tag{2.17}$$

$$a_i = \langle \mathcal{M}, \Psi_i \rangle_w. \tag{2.18}$$

In practice one has to truncate the infinite series in Eq. (2.17) and approximate it by a finite number of summands. One therefore considers the projection of $\mathcal{M}$ onto the subspace $\mathbb{P}_P = \mathrm{span}(\{\Psi_i\}_{i \leq P})$ spanned by the first $P \in \mathbb{N}_{>0}$ basis vectors $\{\Psi_i\}_{i \leq P}$. This projection is given as

$$\mathcal{M}_P(\boldsymbol{x}) = \sum_{i=1}^{P} a_i \Psi_i(\boldsymbol{x}). \tag{2.19}$$

The truncation error or residual $r_P(\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}) - \mathcal{M}_P(\boldsymbol{x}) = \sum_{i=P+1}^{\infty} a_i \Psi_i(\boldsymbol{x})$ is orthogonal with respect to the subspace $\mathbb{P}_P \subset L^2_w(\mathcal{D}_{\boldsymbol{x}})$. This means that $\langle r_P, u_P \rangle_w = \langle u_P, r_P \rangle_w = 0$ for all $u_P \in \mathbb{P}_P$ which is denoted as $r_P \perp \mathbb{P}_P$. One can characterize the residual in the respective Hilbert space norm. The norm $\|r_P\|_w$, or equivalently its square $\|r_P\|^2_w = \sum_{i=P+1}^{\infty} a_i^2$, is minimized over the subspace $\mathbb{P}_P$ and converges to zero for $P \to \infty$. To make this explicit we write

$$\|r_P\|^2_w = \inf_{u_P \in \mathbb{P}_P} \|\mathcal{M} - u_P\|^2_w, \quad \lim_{P \to \infty} \|r_P\|^2_w = 0. \tag{2.20}$$

A geometric interpretation of the subspace projection in Eq. (2.19) is provided in the familiar-looking Fig. 2.4. The model $\mathcal{M} = a_1\Psi_1 + a_2\Psi_2 + a_3\Psi_3$ therein is an element of a three-dimensional function space $\mathbb{P}_3 = \mathrm{span}(\{\Psi_1, \Psi_2, \Psi_3\})$. Its projection onto the two-dimensional subspace $\mathbb{P}_2 = \mathrm{span}(\{\Psi_1, \Psi_2\})$ is $\mathcal{M}_2 = a_1\Psi_1 + a_2\Psi_2$. The residual is simply the difference $r_2 = \mathcal{M} - \mathcal{M}_2 = a_3\Psi_3$ between the true function and its approximation. One has the orthogonality $r_2 \perp \mathbb{P}_2$.

In uncertainty analysis one often equates the positive weight function $w(\boldsymbol{x}) = \pi(\boldsymbol{x})$ with the probability density of the inputs in Eq. (2.2). This adds a probabilistic interpretation. Elements $u, v \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ define random variables $u(\boldsymbol{X})$ and $v(\boldsymbol{X})$ with finite variance. The inner product in Eq. (2.16) is an expectation value in the sense that $\langle u, v \rangle_\pi = \mathbb{E}[u(\boldsymbol{X})v(\boldsymbol{X})]$. By analogy with Eq. (2.17), the output in Eq. (2.8) can be expanded in terms of basis random variables $\{\Psi_i(\boldsymbol{X})\}_{i \in \mathbb{N}_{>0}}$ as $\mathcal{M}(\boldsymbol{X}) = \sum_{i=1}^{\infty} a_i \Psi_i(\boldsymbol{X})$. This is called a *stochastic spectral expansion* at times.

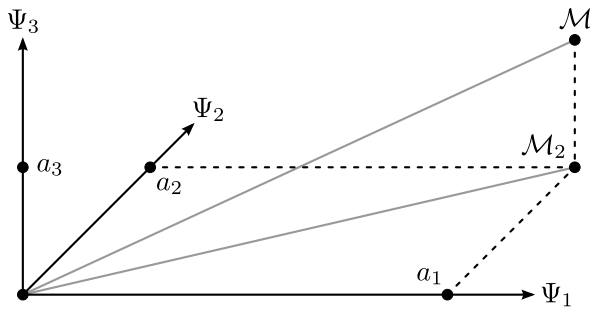


Figure 2.4: Orthogonal projection.

2.3.2 Polynomial approximations

Now we assume that the random input variables are independent, i.e. their density factorizes into $\pi(\boldsymbol{x}) = \pi_1(x_1)\dots\pi_M(x_M)$. The space $L^2_\pi(\mathcal{D}_{\boldsymbol{x}}) = \{u\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R} | \mathbb{E}[u^2(\boldsymbol{X})] < \infty\} \cong L^2_{\pi_1}(\mathcal{D}_{x_1}) \otimes \dots \otimes L^2_{\pi_M}(\mathcal{D}_{x_M})$ is then isomorphic to the tensor product of the Hilbert spaces $L^2_{\pi_i}(\mathcal{D}_{x_i}) = \{u_i\colon \mathcal{D}_{x_i} \to \mathbb{R} | \mathbb{E}[u_i^2(X_i)] < \infty\}$ for $i = 1,\dots,M$. They have the inner products $\langle u_i, v_i\rangle_{\pi_i} = \mathbb{E}[u_i(X_i)v_i(X_i)]$ for $u_i, v_i \in L^2_{\pi_i}(\mathcal{D}_{x_i})$. For two elements $u = u_1 \otimes \dots \otimes u_M$ and $v = v_1 \otimes \dots \otimes v_M$ of the Hilbert space tensor product one thus has $\langle u, v\rangle_\pi = \langle u_1, v_1\rangle_{\pi_1} \dots \langle u_M, v_M\rangle_{\pi_M}$. If the spaces $L^2_{\pi_i}(\mathcal{D}_{x_i})$ have bases $\{\psi^{(i)}_{\alpha_i}\}_{\alpha_i\in\mathbb{N}}$ then $\{\psi^{(1)}_{\alpha_1} \otimes \dots \otimes \psi^{(M)}_{\alpha_M}\}_{\alpha_1,\dots,\alpha_M\in\mathbb{N}}$ forms a basis of $L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$.

Before constructing a basis of $L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ this way, convenient bases of the $L^2_{\pi_i}(\mathcal{D}_{x_i})$ function spaces are discussed. Polynomials that are orthogonal with respect to inner products whose weight function corresponds to common probability densities often constitute such bases. Those polynomials are intimately related to the distributional moments. One starts by considering a family of polynomials $\{\psi^{(i)}_{\alpha_i}(x_i)\}_{\alpha_i\in\mathbb{N}}$ in a single input variable $x_i \in \mathcal{D}_{x_i}$. Here, $\alpha_i \in \mathbb{N}$ is the polynomial degree. The orthogonality relation is $\langle \psi^{(i)}_{\alpha_i}, \psi^{(i)}_{\beta_i}\rangle_{\pi_i} = \delta_{\alpha_i\beta_i}\|\psi^{(i)}_{\alpha_i}\|^2_{\pi_i}$. Four well-known distributions and their associated orthogonal polynomials are listed in Table 2.1. The uniform distribution is linked to the Legendre polynomials. Their first six members $\{P_\alpha(x)\}_{\alpha=0}^5$ up to degree five in a single variable $x \in [-1, 1]$ are shown in Fig. 2.5.

Table 2.1: Orthogonal polynomials.

Distribution	Support	Polynomials
Gaussian	$(-\infty, \infty)$	Hermite
Uniform	$[-1, 1]$	Legendre
Beta	$[-1, 1]$	Jacobi
Gamma	$[0, \infty)$	Laguerre

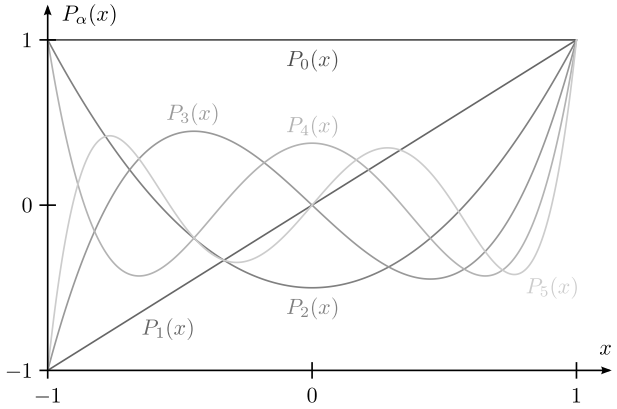


Figure 2.5: Legendre polynomials.

Such polynomials often constitute a complete set in $L^2_{\pi_i}(\mathcal{D}_{x_i})$ [72–74]. A notable exception is the lognormal distribution for which one has to take care of some subtleties related to moment determinacy [75]. This distribution is not uniquely determined by its sequence of moments, i.e. other genuinely different distributions feature the same moments. The corresponding polynomials are not dense in the full space of mean-square integrable functions. They form a set of elements that is orthogonal but not complete, i.e. $\mathrm{span}(\{\psi^{(i)}_{\alpha_i}\}_{\alpha_i\in\mathbb{N}}) \subsetneq L^2_{\pi_i}(\mathcal{D}_{x_i})$. Intuitively this becomes clearer by realizing that the linear hull of the polynomials is at most the intersection of different function spaces in which the polynomials are orthogonal.

In this situation one has to consider a slightly less general case, e.g. mean-square integrability is replaced by the stronger assumption that the model is actually in the subspace spanned by the polynomials. One may also argue that only the model's projection onto the respective subspace is considered in this case, i.e. the orthogonal complement is simply discarded. Similar arguments can be evoked when constructing a finite number of polynomials that are orthogonal with respect to a distribution that is empirically known by its first few

moments only [76–78]. Another obvious possibility to bypass the problem is to transform the lognormal to a normal distribution by taking the natural logarithm.

Speaking of parameter transformations, it may very well be necessary to reparametrize the problem either way. This happens so as to guarantee that the inputs have convenient bounds and probability distributions, e.g. the standard types that are compiled in Table 2.1. A related discussion on probabilistic model parametrizations is found in Section 3.5. Uniform distributions with arbitrary bounds can be linearly transformed into $\pi_i(x_i) = \mathcal{U}(x_i|-1,1)$ with $\mathcal{U}(x_i|-1,1) = 1/2$ for $x_i \in [-1,1]$ and $\mathcal{U}(x_i|-1,1) = 0$ otherwise. Similarly, Gaussians with arbitrary mean and variance can be shifted and scaled into a standard normal $\pi_i(x_i) = \mathcal{N}(x_i|0,1) = \exp(-x_i^2/2)/\sqrt{2\pi}$ with mean $\mu_{X_i} = 0$ and variance $\sigma_{X_i}^2 = 1$. We assume that our parameters are already of such a standard form.

Having univariate polynomials $\{\psi_{\alpha_i}^{(i)}(x_i)\}_{\alpha_i \in \mathbb{N}}$ for each variable x_i, multivariate polynomials in $\boldsymbol{x}$ are constructed via tensorization. We introduce a multi-index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M) \in \mathbb{N}^M$ for bookkeeping purposes. Then a set of multivariate polynomials $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\}_{\boldsymbol{\alpha} \in \mathbb{N}^M}$ is given through

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \prod_{i=1}^{M} \psi_{\alpha_i}^{(i)}(x_i). \tag{2.21}$$

In this construction, the orthogonality of the multivariate polynomials carries over from the univariate ones. This is verified by $\langle \psi_{\boldsymbol{\alpha}}, \psi_{\boldsymbol{\beta}} \rangle_\pi = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}} \|\psi_{\boldsymbol{\alpha}}\|_\pi^2 = \delta_{\alpha_1\beta_1} \dots \delta_{\alpha_M\beta_M} \|\psi_{\alpha_1}^{(1)}\|_{\pi_1}^2 \dots \|\psi_{\alpha_M}^{(M)}\|_{\pi_M}^2$.

The polynomials defined in Eq. (2.21) are complete in $L_\pi^2(\mathcal{D}_{\boldsymbol{x}})$. In the vein of Eqs. (2.17) and (2.18), the expansion of the model with respect to the multivariate polynomial basis $\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\}_{\boldsymbol{\alpha} \in \mathbb{N}^M}$ is given as

$$\mathcal{M} = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} a_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}, \tag{2.22}$$

$$a_{\boldsymbol{\alpha}} = \langle \mathcal{M}, \psi_{\boldsymbol{\alpha}} \rangle_\pi / \|\psi_{\boldsymbol{\alpha}}\|_\pi^2. \tag{2.23}$$

The stochastic spectral version of Eqs. (2.22) and (2.23) is a certain *polynomial chaos expansion* (PCE) $\tilde{Y} = \mathcal{M}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} a_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{X})$ of the random variable in Eq. (2.8). Note that one can normalize the basis elements $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^M}$ and linearly order them according to their multi-index $\boldsymbol{\alpha}$.

An ordering scheme of some type can also facilitate the truncation of the series in Eq. (2.22) as in Eq. (2.19). In a quite natural way one can impose restrictions on the total degree $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^M |\alpha_i|$ of the polynomials in Eq. (2.21). A finite number of terms is obtained by keeping only the ones with a total degree equal to or smaller than a certain $p \in \mathbb{N}$. These are the terms indexed by $\boldsymbol{\alpha} \in \mathcal{A}_p$ in $\mathcal{A}_p = \{\boldsymbol{\alpha} \in \mathbb{N}^M \colon \|\boldsymbol{\alpha}\|_1 \leq p\}$. The cardinality P of this set is dependent on the input dimensionality M and the maximal degree p through

$$P = \binom{M+p}{p} = \frac{(M+p)!}{M!\,p!}. \tag{2.24}$$

The fast growth of the total number of terms P in Eq. (2.24) can be ascribed to the curse of dimensionality. That issue is further discussed in Section 2.6. By retaining only the terms $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}_p}$ in the expansion the best polynomial approximation of total degree p is given by

$$\mathcal{M}_P = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_p} a_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}. \tag{2.25}$$

Using a polynomial basis has some appealing advantages. In the first place it is easy to understand and interpret, e.g. the contributing terms can be classified according to their polynomial degrees and multivariate structure. One can distinguish between low-order and high-order terms or identify individual contributions and mutual interactions of input variables. Furthermore, one could argue that many physical laws and models lend themselves to polynomial approximations, e.g. think about Taylor expansions as in Section 2.2.

For intents of uncertainty quantification it is convenient that the coefficients of polynomial expansions are related to the statistical moments in Eqs. (2.9) and (2.10). One has for instance

$$\mu_{\tilde{Y}} = a_{\boldsymbol{0}} \|\psi_{\boldsymbol{0}}\|_\pi, \quad \sigma_{\tilde{Y}}^2 = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M \setminus \{\boldsymbol{0}\}} a_{\boldsymbol{\alpha}}^2 \|\psi_{\boldsymbol{\alpha}}\|_\pi^2. \tag{2.26}$$

The constant expansion term determines the expected value $\mu_{\tilde{Y}} = \mathbb{E}[\mathcal{M}(\boldsymbol{X})]$ in Eq. (2.26). So do the remaining terms for the variance $\sigma_{\tilde{Y}}^2 = \mathrm{Var}[\mathcal{M}(\boldsymbol{X})]$.

2.4 Discrete least squares

After merely representing the simulator in terms of basis functions in Eqs. (2.17) and (2.22), one has to actually compute the expansion coefficients in Eqs. (2.18) and (2.23). A straightforward approach is to perform the involved integrals explicitly. We discuss an alternative method that is based on the characterization of the subspace projections in Eqs. (2.19) and (2.25) as the minimizers of the residual norm in Eq. (2.20). In order to highlight the analogy of the formulations, this continuous least squares property is revisited before its natural discretization is discussed. A linear least squares minimization problem [79] arises that way.

Let us assume that we have a set of linearly independent basis functions $\{\psi_j(\boldsymbol{x})\}_{j\leq P}$, i.e. no basis function is a linear combination of the others. They are used in the ansatz $\mathcal{M}_P(\boldsymbol{x}) = \sum_{j=1}^{P} a_j\psi_j(\boldsymbol{x}) \in \operatorname{span}(\{\psi_j(\boldsymbol{x})\}_{j\leq P})$ for finding the best approximation of the simulator $\mathcal{M}(\boldsymbol{x})$ in that it minimizes the residual. To this effect, the coefficient vector $\boldsymbol{a} = (a_1, \dots, a_P)^\top$ is chosen such that

$$\boldsymbol{a} = \underset{\boldsymbol{a}^\star\in\mathbb{R}^P}{\arg\min} \int_{\mathcal{D}_{\boldsymbol{x}}} \left(\mathcal{M}(\boldsymbol{x}) - \sum_{j=1}^{P} a_j^\star\psi_j(\boldsymbol{x})\right)^2 \pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{2.27}$$

The unique minimum of this *continuous least squares* problem is then obtained when the partial derivatives $\partial\|\mathcal{M} - \mathcal{M}_P\|_\pi^2/\partial a_{j'} = 0$ are zero for $j' = 1, \dots, P$. This results in the *continuous normal equations*

$$\sum_{j=1}^{P} a_j \int_{\mathcal{D}_{\boldsymbol{x}}} \psi_j(\boldsymbol{x})\psi_{j'}(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{M}(\boldsymbol{x})\psi_{j'}(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{2.28}$$

If one chooses orthogonal basis functions, the system of equations can be easily solved for the coefficients $a_{j'}$. This exactly yields the orthogonal projections $a_{j'} = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{M}(\boldsymbol{x})\psi_{j'}(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} / \int_{\mathcal{D}_{\boldsymbol{x}}} \psi_{j'}^2(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

A discretization of the continuous case is based on a finite number $K \geq P$ of model runs. They are performed for a representative sample of input values that is called the *experimental design*

$$\mathcal{X} = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(K)}). \tag{2.29}$$

We consider the scenario that these inputs are independently sampled from the input probability distribution, i.e. $\pi_{\mathcal{X}}(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(K)}) = \pi(\boldsymbol{x}^{(1)}) \dots \pi(\boldsymbol{x}^{(K)})$. Afterwards one has to compute the corresponding responses $\mathcal{Y} = (\mathcal{M}(\boldsymbol{x}^{(1)}), \dots, \mathcal{M}(\boldsymbol{x}^{(K)}))^\top$. The *design matrix* $\boldsymbol{A} \in \mathbb{R}^{K\times P}$ is composed as

$$\boldsymbol{A} = \begin{pmatrix} 1 & \psi_2(\boldsymbol{x}^{(1)}) & \dots & \psi_P(\boldsymbol{x}^{(1)}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \psi_2(\boldsymbol{x}^{(K)}) & \dots & \psi_P(\boldsymbol{x}^{(K)}) \end{pmatrix}. \tag{2.30}$$

For $k = 1, \dots, K$ and $l = 1, \dots, P$ this Vandermonde-like matrix has the entries $A_{k,l} = \Psi_l(\boldsymbol{x}^{(k)})$. Since one typically has a constant term with $\psi_1(\boldsymbol{x}) = 1$, the elements of the first column of the design matrix in Eq. (2.30) are equal to one. For the computed model outputs one can now establish the equations

$$\begin{pmatrix} \mathcal{M}(\boldsymbol{x}^{(1)}) \\ \vdots \\ \mathcal{M}(\boldsymbol{x}^{(K)}) \end{pmatrix} = \begin{pmatrix} 1 & \psi_2(\boldsymbol{x}^{(1)}) & \dots & \psi_P(\boldsymbol{x}^{(1)}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \psi_2(\boldsymbol{x}^{(K)}) & \dots & \psi_P(\boldsymbol{x}^{(K)}) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{pmatrix} + \begin{pmatrix} r_P(\boldsymbol{x}^{(1)}) \\ \vdots \\ r_P(\boldsymbol{x}^{(K)}) \end{pmatrix}. \tag{2.31}$$

In more compact notation one can write Eq. (2.31) as $\mathcal{Y} = \boldsymbol{A}\boldsymbol{a} + \boldsymbol{r}_P$. The vector $\boldsymbol{r}_P = (r_P(\boldsymbol{x}^{(1)}), \dots, r_P(\boldsymbol{x}^{(K)}))^\top$ gathers the differences $r_P(\boldsymbol{x}^{(k)}) = \mathcal{M}(\boldsymbol{x}^{(k)}) - \mathcal{M}_P(\boldsymbol{x}^{(k)})$ between the model and its best approximation over the whole input domain.

It is now tempting to ask for the $\hat{\mathcal{M}}_P(\boldsymbol{x}) = \sum_{j=1}^{P} \hat{a}_j\psi_j(\boldsymbol{x}) \in \operatorname{span}(\{\psi_j(\boldsymbol{x})\}_{j\leq P})$ that best approximates the response data $\mathcal{Y}$ over the discrete points in the experimental design $\mathcal{X}$ rather than the full space $\mathcal{D}_{\boldsymbol{x}}$. Therefore consider the system $\mathcal{Y} = \boldsymbol{A}\hat{\boldsymbol{a}} + \boldsymbol{r}_{P,K}$, where the residual vector $\boldsymbol{r}_{P,K} = (r_{P,K}(\boldsymbol{x}^{(1)}), \dots, r_{P,K}(\boldsymbol{x}^{(K)}))^\top$ collects the differences $r_{P,K}(\boldsymbol{x}^{(k)}) = \mathcal{M}(\boldsymbol{x}^{(k)}) - \hat{\mathcal{M}}_P(\boldsymbol{x}^{(k)})$. The coefficients $\hat{\boldsymbol{a}} = (\hat{a}_1, \dots, \hat{a}_P)^\top$ are chosen such that $\|\boldsymbol{r}_{P,K}\|_2^2 = r_{P,K}^2(\boldsymbol{x}^{(1)}) + \dots + r_{P,K}^2(\boldsymbol{x}^{(K)})$ is minimal. That describes the *discrete least squares* problem

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}^\star\in\mathbb{R}^P}{\arg\min} \|\mathcal{Y} - \boldsymbol{A}\boldsymbol{a}^\star\|_2^2. \tag{2.32}$$

This is the discrete companion of Eq. (2.27). Just as Eq. (2.28) was obtained, we zero the partial derivatives $\partial\|\mathcal{Y} - \boldsymbol{A}\hat{\boldsymbol{a}}\|_2^2/\partial\hat{a}_{j'} = 0$ for $j' = 1, \dots, P$ in order to derive the *discrete normal equations*

$$\boldsymbol{A}^\top \boldsymbol{A}\hat{\boldsymbol{a}} = \boldsymbol{A}^\top \mathcal{Y}. \tag{2.33}$$

If one assumes that the columns of $\boldsymbol{A}$ are linearly independent, i.e. $\boldsymbol{A}^\top\boldsymbol{A}$ is positive-definite and therefore invertible, then the unique *ordinary least squares* (OLS) solution is

$$\hat{\boldsymbol{a}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \mathcal{Y}. \tag{2.34}$$

Notice that one can write the gradient conditions in Eq. (2.33) as $\boldsymbol{A}^\top(\mathcal{Y} - \boldsymbol{A}\hat{\boldsymbol{a}}) = \boldsymbol{0}$. Geometrically interpreted this means that $\boldsymbol{r}_{P,K} \perp \mathrm{col}(\boldsymbol{A})$, i.e. the residual vector $\boldsymbol{r}_{P,K} = \mathcal{Y} - \boldsymbol{A}\hat{\boldsymbol{a}}$ is orthogonal to the column space $\mathrm{col}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{a}^\star \colon \boldsymbol{a}^\star \in \mathbb{R}^P\}$ of the design matrix $\boldsymbol{A}$. Sometimes one defines the *hat matrix* $\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$. It projects $\mathcal{Y}$ onto the column space by $\hat{\mathcal{Y}} = \boldsymbol{H}\mathcal{Y} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathcal{Y} = \boldsymbol{A}\hat{\boldsymbol{a}}$, where $\hat{\boldsymbol{a}}$ is from Eq. (2.34).

The abovementioned projection gives indeed just the approximations $\hat{\mathcal{Y}} = (\hat{\mathcal{M}}_P(\boldsymbol{x}^{(1)}), \dots, \hat{\mathcal{M}}_P(\boldsymbol{x}^{(K)}))^\top$ of how the model responds to the experimental design $\mathcal{X}$. Yet, the exact values are known anyway. For arbitrary inputs values $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ the emulation of the original simulator $\mathcal{M}(\boldsymbol{x})$ is based on

$$\hat{\mathcal{M}}_P(\boldsymbol{x}) = \sum_{j=1}^{P} \hat{a}_j \psi_j(\boldsymbol{x}). \tag{2.35}$$

2.4.1 Prediction errors

At this point one ought to give particular attention to various kinds of prediction errors [80, 81]. Notwithstanding that minimizing the empirical residual norm in Eq. (2.32) is a plausible way of proceeding, it does not necessarily warrant accuracy of the metamodel in Eq. (2.35) for unseen input data. There are certainly various ways of quantifying the predictions errors. A short overview of a few related concepts is provided now. Since the terminology is inconsistent throughout the literature in various fields, one always has to specify the error measure under consideration clearly.

Let us denote the surrogate model that was obtained for a specific experimental design $\mathcal{X}$ by $\hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x})$. The *generalization error* of this predictor is defined as the expectation value under the input distribution

$$\mathrm{Err}\left[\hat{\mathcal{M}}_P^{\mathcal{X}}\right] = \mathbb{E}\left[\left(\mathcal{M}(\boldsymbol{X}) - \hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{X})\right)^2\right] = \int_{\mathcal{D}_{\boldsymbol{x}}} \left(\mathcal{M}(\boldsymbol{x}) - \hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x})\right)^2 \pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{2.36}$$

This is exactly the error that we tried to minimize. One may define the *expected generalization error* by additionally averaging over the distribution of the experimental design

$$\overline{\mathrm{Err}}\left[\hat{\mathcal{M}}_P\right] = \mathbb{E}_{\mathcal{X}}\left[\mathrm{Err}\left[\hat{\mathcal{M}}_P^{\mathcal{X}}\right]\right]. \tag{2.37}$$

This can be interpreted as an error of the procedure or rule of computing a predictor rather than of a specific predictor itself. One can also define the *expected prediction error* at a single point $\boldsymbol{x}_0 \in \mathcal{D}_{\boldsymbol{x}}$ as

$$\overline{\mathrm{Err}}\left[\hat{\mathcal{M}}_P(\boldsymbol{x}_0)\right] = \mathbb{E}_{\mathcal{X}}\left[\left(\mathcal{M}(\boldsymbol{x}_0) - \hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x}_0)\right)^2\right] = \mathrm{Bias}_{\mathcal{X}}^2\left[\hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x}_0)\right] + \mathrm{Var}_{\mathcal{X}}\left[\hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x}_0)\right]. \tag{2.38}$$

The classical trade-off between the estimation bias $\mathrm{Bias}_{\mathcal{X}}[\hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x}_0)] = \mathbb{E}_{\mathcal{X}}[\hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x}_0)] - \mathcal{M}(\boldsymbol{x}_0)$ and the variance $\mathrm{Var}_{\mathcal{X}}[\hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x}_0)] = \mathbb{E}_{\mathcal{X}}[(\hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x}_0) - \mathbb{E}_{\mathcal{X}}[\hat{\mathcal{M}}_P^{\mathcal{X}}(\boldsymbol{x}_0)])^2]$ has emerged here.

The generalization error in Eq. (2.36) measures the inaccuracy of a single approximation in a weighted sense over the model input space. This is a quantity we would like to know once we have a metamodel at hand. In contrast, the errors in Eqs. (2.37) and (2.38) provide statistical information about the approximation method with respect to the sampling distribution of the experimental design. These are quantities whose minimization is targeted in the design of efficient metamodeling algorithms.

After the computation of several metamodels, e.g. of various orders and with different experimental designs, the ones that generalize poorly can be separated out in a validation step. This usually requires additional simulations which the metamodeling predictions can be checked against. In Chapter 8 a discussion of leave-one-out cross validation in the present linear context is found. It allows one to cross-validate the metamodeling procedure against points in the experimental design without further simulations. This can guide the practical computation and selection of such metamodels that satisfactorily generalize beyond the experimental design.

2.4.2 Relation to linear regression

It is interesting to note that the residual error minimization for function approximation is formally evocative of statistical linear regression [82]. These two problems are different in nature, though. The function approximation problem deals with nonparametric approximations to completely unknown and possibly complex functions. It features an experimental design with random inputs and noise-free simulations. On the contrary, classical regression is concerned with simple models that are assumed to represent the truth and have a well-known parametric form. It is a statistical estimation problem involving fixed covariates and noisy observations. The linear statistical model is discussed in Section 3.6.2 later on. Under some standard assumptions related to the measurement errors, the maximum likelihood estimate indeed coincides with the least squares minimizer. Hence, the extremization objective and its solution are technically identical for both the function approximation and the parameter estimation problem.

However, other results from linear regression theory, e.g. statements regarding the unbiasedness of the estimator of the coefficient vector, cannot be transferred. Since this is often neglected, the latter thought shall be engrossed a bit. The bias of a statistical estimator is the difference between the expectation value of this estimator under its sampling distribution and the true value. Let us denote the expectation value of the OLS solution in Eq. (2.34) over the population distribution of the experimental design by $\mathbb{E}_{\mathcal{X}}[\hat{\boldsymbol{a}}]$. With Eq. (2.31) we immediately see that

$$\mathbb{E}_{\mathcal{X}}[\hat{\boldsymbol{a}}^{\mathcal{X}}] = \mathbb{E}_{\mathcal{X}}[(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathcal{Y}] = \mathbb{E}_{\mathcal{X}}[(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{a}+\boldsymbol{r}_P)] = \boldsymbol{a} + \mathbb{E}_{\mathcal{X}}[(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{r}_P]. \tag{2.39}$$

Here, all quantities but the true projection coefficient $\boldsymbol{a}$ depend on the design over which the expectation is taken. The corresponding index has been dropped for notational convenience. In the context of function approximation with a random experimental design, the bias of the OLS estimate $\hat{\boldsymbol{a}}$ is thus given as $\mathrm{Bias}_{\mathcal{X}}[\hat{\boldsymbol{a}}] = \mathbb{E}_{\mathcal{X}}[\hat{\boldsymbol{a}}] - \boldsymbol{a} = \mathbb{E}_{\mathcal{X}}[(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{r}_P]$. It is dependent on the vector of residual errors $\boldsymbol{r}_P = \mathcal{Y} - \boldsymbol{A}\boldsymbol{a}$ which is not quite the same as $\boldsymbol{r}_{P,K} = \mathcal{Y} - \boldsymbol{A}\hat{\boldsymbol{a}}$. Thus the bias can only vanish asymptotically in the limit $P \to \infty$ such that $\mathbb{E}[r_P^2(\boldsymbol{X})] \to 0$. For finite $P < \infty$ it is zero only if it happens that $\mathcal{M}(\boldsymbol{x}) = \mathcal{M}_P(\boldsymbol{x}) = \sum_{j=1}^{P} a_j\psi_j(\boldsymbol{x}) \in \mathrm{span}(\{\psi_j(\boldsymbol{x})\}_{j\leq P})$ for which one has $\boldsymbol{r}_P = \boldsymbol{0}$.

2.4.3 Relation to Monte Carlo integration

The matrix $\boldsymbol{A}^\top\boldsymbol{A}$ in Eqs. (2.33) and (2.34) plays a crucial role in least squares regression and its statistical design [83, 84]. It is the symmetric and positive-semidefinite *Gramian matrix* of the columns of the design matrix $\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_P)$. For $j, j' = 1, \ldots, P$ the entries of the Gramian are $\boldsymbol{A}_j^\top\boldsymbol{A}_{j'}$.

If the columns of the design matrix are empirically orthogonal in the sense that for all $j \neq j'$ one has $\boldsymbol{A}_j^\top\boldsymbol{A}_{j'} = \sum_{k=1}^{K}\psi_j(\boldsymbol{x}^{(k)})\psi_{j'}(\boldsymbol{x}^{(k)}) = 0$, then the Gramian matrix becomes diagonal

$$\boldsymbol{A}^\top\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_1^\top\boldsymbol{A}_1 & \ldots & \boldsymbol{A}_1^\top\boldsymbol{A}_P \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_P^\top\boldsymbol{A}_1 & \ldots & \boldsymbol{A}_P^\top\boldsymbol{A}_P \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{K}\psi_1^2(\boldsymbol{x}^{(k)}) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sum_{k=1}^{K}\psi_P^2(\boldsymbol{x}^{(k)}) \end{pmatrix}. \tag{2.40}$$

For the components of $\hat{\boldsymbol{a}} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathcal{Y}$ one then finds $\hat{a}_j = \sum_{k=1}^{K}\psi_j(\boldsymbol{x}^{(k)})\mathcal{M}(\boldsymbol{x}^{(k)}) / \sum_{k=1}^{K}\psi_j^2(\boldsymbol{x}^{(k)})$ for $j = 1, \ldots, P$. This means that the coefficients are estimated independently from each other, in the sense that the estimate $\hat{a}_j$ for a certain j does not at all depend on the inclusion of further terms $\psi_{j'}(\boldsymbol{x})$ with $j' \neq j$.

Moreover, in the case that the experimental design is randomized, this perfectly corresponds to the MC estimate of the orthogonal projection. Even for finite experimental designs with $K < \infty$, a relation between least squares minimization and numerical integration is thus established in the event that empirical orthogonality is fulfilled. In the more general case the two procedures yield different results.

2.5 Multivariate output

Now we consider vector-valued models $\mathcal{M}\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R}^N$ that map inputs $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ to multiple outputs $\tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{x}) \in \mathbb{R}^N$ with $N \in \mathbb{N}_{>0}$. Each component of the output vector $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \ldots, \tilde{y}_N)^\top$ is predicted as $\tilde{y}_i = \mathcal{M}_{\tilde{y}_i}(\boldsymbol{x}) \in \mathbb{R}$ for $i = 1, \ldots, N$ by a function of all inputs denoted as $\mathcal{M}_{\tilde{y}_i}\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R}$. In explicit notation this can be written as

$$\tilde{\boldsymbol{y}} = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_N \end{pmatrix} = \begin{pmatrix} \mathcal{M}_{\tilde{y}_1}(\boldsymbol{x}) \\ \vdots \\ \mathcal{M}_{\tilde{y}_N}(\boldsymbol{x}) \end{pmatrix}. \tag{2.41}$$

One should distinguish between the predicted output $\tilde{\boldsymbol{y}} \in \mathbb{R}^N$ as an element in the vector space $\mathbb{R}^N$ and the predicting model $\mathcal{M}_{\tilde{y}_i} \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ which is assumed to be an element of the function space $L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$. Accordingly, for $i = 1, \ldots, N$ one can represent, project and approximate each $\mathcal{M}_{\tilde{y}_i}$ separately with the theory and methods previously discussed.

One can also represent the model outputs $\tilde{\boldsymbol{y}} \in \mathbb{R}^N$ with respect to another basis of the model output space. Let $\{\boldsymbol{\phi}_i\}_{i=1}^N$ be an orthonormal basis of the Euclidean vector pace $\mathbb{R}^N$ which is different from the standard reference system $\{\boldsymbol{e}_i\}_{i=1}^N$. As an alternative to the naturally suggested representation $\tilde{\boldsymbol{y}} = \sum_{i=1}^N \tilde{y}_i \boldsymbol{e}_i$, the model responses can be expanded as

$$\tilde{\boldsymbol{y}} = \sum_{i=1}^{N} \tilde{z}_i \boldsymbol{\phi}_i. \tag{2.42}$$

Here, $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \ldots, \tilde{z}_N)^\top$ is the coordinate vector of the output relative to the alternative basis, where each coordinate is given as $\tilde{z}_i = \tilde{\boldsymbol{y}}^\top \boldsymbol{\phi}_i = \tilde{\boldsymbol{y}} \cdot \boldsymbol{\phi}_i$ for $i = 1, \ldots, N$. As a function of the uncertain model parameters, each $\tilde{z}_i = \mathcal{M}_{\tilde{z}_i}(\boldsymbol{x})$ is predicted by the corresponding scalar-valued function $\mathcal{M}_{\tilde{z}_i} \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$.

Thus far, the canonical representation in Eq. (2.41) has been merely reformulated in Eq. (2.42). Even though the two formulations before and after the change a basis are technically equivalent, they may differ in their eligibility for signal compaction [85, 86]. If the output dimension is high, it is inconvenient at the very least to have to treat multiple outputs individually. Moreover, different model outputs are often redundant to some degree anyhow. This motivates to consider bases where the essential features of the model are captured by just a few dominant terms. Such considerations are especially important against the backdrop of the following section.

2.6 Curse of dimensionality

High-dimensionality forms an obstacle to the analysis of many complex systems. That is widely agreed, even though the degree as to which it applies is strongly dependent on the academic field. The so-called *curse of dimensionality* [87, 88] very generally refers to difficulties in the characterization of high-dimensional objects with discrete information. Various different manifestations of this problem are related to the volume and geometry of high-dimensional spaces. They play a major role in much of machine learning [89, 90], learning theory [91, 92] and multivariate statistics [93, 94].

Examples that are often used to illustrate the immense volume and the counterintuitive behavior of Euclidean norms and distances in high-dimensional spaces include the following. The volume of a hypersphere with fixed radius, e.g. a unit hypersphere embedded into a unit hypercube whose volume is always one, drops to zero. Most of the volume of a high-dimensional sphere is located in a thin shell underneath its surface, rather than in the interior. Similarly, most of the probability mass of a multivariate Gaussian distribution concentrates around a shell distant from the expected value, rather than in the bell. In one, two and three space dimensions, data points on a regular grid occupying a unit interval, square and cube are shown in Fig. 2.6, respectively. In order to ensure a consistent coverage of the space, the sample size has to grow exponentially with the dimensionality. Vice versa, uniform data with a fixed sample size quickly become isolated.

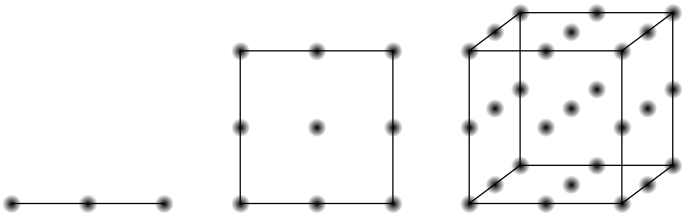


Figure 2.6: Curse of dimensionality.

The empty space phenomenon immediately shows up in UQ problems for the experimental design in Eq. (2.29). Training sets with a fixed sample size fail to representatively cover the input space. A sufficient coverage would require an exploding number of training runs and is therefore computationally infeasible. The curse of dimensionality manifests also in Eq. (2.24) which obstructs the expansion of high-dimensional functions in terms of high-order polynomials. Yet another instantiation of the problem is in Eq. (2.41) where all high-dimensional model outputs must be considered individually.

Confronted with this seemingly hopeless situation, one may contemplate surrender. But fortunately there is the blessing of *sparsity*, i.e. real-world problems in high-dimensional ambient spaces often feature characteristic low-dimensional sub-structures that are hidden at the first sight. By uncovering such essential patterns of

information one can break or at least try to smother the curse. Given a real data set with many features, techniques for data dimensionality reduction can be applied [95]. Given an engineering model, a subset of important directions in the multi-dimensional input space can be identified [96]. Another type of sparsity can be sought after directly in the representation of a signal [97]. A hypothetical basis that already contains the actual signal would precisely require that one term. Even though this case is improbable, sparse recovery methods can be applied in contexts where a favorable basis is known.

The solution of complex engineering problems in the UQ domain often requires to use a combination of methods for tackling high-dimensionality at the various modeling levels. An example can be found in Chapter 10 wherein a hydrological simulator with a collapsed input space will be analyzed. Also the dimension of the simulator response will be reduced. After that a regularized least squares problem will be solved that seeks sparsity in the functional basis expansion.

References

[1] N. Metropolis and S. Ulam. "The Monte Carlo Method". In: *Journal of the American Statistical Association* 44.247 (1949), pp. 335–341. DOI: 10.1080/01621459.1949.10483310.

[2] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45. DOI: 10.1115/1.3662552.

[3] R. E. Kalman and R. S. Bucy. "New Results in Linear Filtering and Prediction Theory". In: *Journal of Basic Engineering* 83.1 (1961), pp. 95–108. DOI: 10.1115/1.3658902.

[4] R. C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Computational Science and Engineering. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2014.

[5] T. J. Sullivan. *Introduction to Uncertainty Quantification*. Texts in Applied Mathematics 63. Cham, Switzerland: Springer International Publishing, 2015. DOI: 10.1007/978-3-319-23395-6.

[6] W. L. Oberkampf and C. J. Roy. *Verification and Validation in Scientific Computing*. Cambridge, UK: Cambridge University Press, 2010. DOI: 10.1017/CBO9780511760396.

[7] D. J. Murray-Smith. *Testing and Validation of Computer Simulation Models: Principles, Methods and Applications*. Simulation Foundations, Methods and Applications. Cham, Switzerland: Springer International Publishing, 2015. DOI: 10.1007/978-3-319-15099-4.

[8] E. de Rocquigny, N. Devictor, and S. Tarantola, eds. *Uncertainty in Industrial Practice: A Guide to Quantitative Uncertainty Management*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2008. DOI: 10.1002/9780470770733.

[9] E. de Rocquigny. *Modelling Under Risk and Uncertainty: An Introduction to Statistical, Phenomenological and Computational Methods*. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2012. DOI: 10.1002/9781119969495.

[10] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. 2nd ed. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2002. DOI: 10.1137/1.9780898718027.

[11] P. Deuflhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing: An Introduction*. 2nd ed. Texts in Applied Mathematics 43. New York: Springer, 2003. DOI: 10.1007/978-0-387-21584-6.

[12] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. 2nd ed. Texts in Applied Mathematics 37. Springer Berlin Heidelberg, 2007. DOI: 10.1007/b98885.

[13] P. Fornasini. *The Uncertainty in Physical Measurements: An Introduction to Data Analysis in the Physics Laboratory*. New York: Springer, 2008. DOI: 10.1007/978-0-387-78650-6.

[14] S. V. Gupta. *Measurement Uncertainties: Physical Parameters and Calibration of Instruments*. Springer-Verlag Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-20989-5.

[15] M. Grabe. *Measurement Uncertainties in Science and Technology*. 2nd ed. Springer-Verlag Berlin Heidelberg, 2014. DOI: 10.1007/978-3-319-04888-8.

[16] B. M. Ayyub and G. J. Klir. *Uncertainty Modeling and Analysis in Engineering and the Sciences*. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2006. DOI: 10.1201/9781420011456.

[17] V. Barnett. *Comparative Statistical Inference*. 3rd ed. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 1999. DOI: 10.1002/9780470316955.

[18] D. R. Cox. *Principles of Statistical Inference*. Cambridge, UK: Cambridge University Press, 2006. DOI: 10.1017/CBO9780511813559.

[19] F. J. Samaniego. *A Comparison of the Bayesian and Frequentist Approaches to Estimation*. Springer Series in Statistics. New York: Springer, 2010. DOI: 10.1007/978-1-4419-5941-6.

[20] A. Der Kiureghian and O. Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009), pp. 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

[21] R. E. Kass. "Statistical Inference: The Big Picture". In: *Statistical Science* 26.1 (2011), pp. 1–9. DOI: 10.1214/10-STS337.

[22] M. Grigoriu. *Stochastic Calculus: Applications in Science and Engineering*. Boston, Massachusetts, USA: Birkhäuser, 2002.

[23] M. Grigoriu. *Stochastic Systems: Uncertainty Quantification and Propagation*. Springer Series in Reliability Engineering. London, UK: Springer-Verlag, 2012. DOI: 10.1007/978-1-4471-2327-9.

[24] M. Lemaire, A. Chateauneuf, and J.-C. Mitteau. *Structural Reliability*. London, UK: ISTE Ltd, 2009. DOI: 10.1002/9780470611708.

[25] A. S. Nowak and K. R. Collins. *Reliability of Structures*. 2nd ed. Boca Raton, Florida, USA: CRC Press, 2013.

[26] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2005. DOI: 10.1137/1.9780898717921.

[27] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Applied Mathematical Sciences 160. New York: Springer, 2005. DOI: 10.1007/b138659.

[28] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge, UK: Cambridge University Press, 2002. DOI: 10.1017/CBO9780511802270.

[29] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. 2nd ed. Springer-Verlag Berlin Heidelberg, 2009. DOI: 10.1007/978-3-642-03711-5.

[30] K.-T. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments*. Computer Science and Data Analysis Series. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2006. DOI: 10.1201/9781420034899.

[31] A. I. J. Forrester, A. Sóbester, and A. J. Keane. *Engineering Design via Surrogate Modelling: A Practical Guide*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2008. DOI: 10.1002/9780470770801.

[32] Z.-Q. Qu. *Model Order Reduction Techniques: With Applications in Finite Element Analysis*. London, UK: Springer-Verlag, 2004. DOI: 10.1007/978-1-4471-3827-3.

[33] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Advances in Design and Control. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2005. DOI: 10.1137/1.9780898718713.

[34] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2004. DOI: 10.1002/0470870958.

[35] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis: The Primer*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2008. DOI: 10.1002/9780470725184.

[36] T. Bedford and R. Cooke. *Probabilistic Risk Analysis: Foundations and Methods*. Cambridge, UK: Cambridge University Press, 2001. DOI: 10.1017/CBO9780511813597.

[37] T. Aven. *Foundations of Risk Analysis*. 2nd ed. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2012. DOI: 10.1002/9781119945482.

[38] J. W. Herrmann. *Engineering Decision Making and Risk Management*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2015.

[39] M. J. Kochenderfer. *Decision Making Under Uncertainty: Theory and Application*. MIT Lincoln Laboratory Series. Cambridge, Massachusetts, USA: The MIT Press, 2015.

[40] S. H. Lee and W. Chen. "A comparative study of uncertainty propagation methods for black-box-type problems". In: *Structural and Multidisciplinary Optimization* 37.3 (2009), pp. 239–253. DOI: 10.1007/s00158-008-0234-7.

[41] M. Arnst and J.-P. Ponthot. "An overview of nonintrusive characterization, propagation, and sensitivity analysis of uncertainties in computational mechanics". In: *International Journal for Uncertainty Quantification* 4.5 (2014), pp. 387–421. DOI: 10.1615/Int.J.UncertaintyQuantification.2014006990.

[42] T. T. Soong. *Fundamentals of Probability and Statistics for Engineers*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2004.

[43] H. Benaroya and S. M. Han. *Probability Models in Engineering and Science*. Mechanical Engineering 193. Boca Raton, Florida, USA: CRC Press, 2005.

[44] H. Schwarzlander. *Probability Concepts and Theory for Engineers*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2011. DOI: 10.1002/9781119990895.

[45] M. Kamiński. *The Stochastic Perturbation Method for Computational Mechanics*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2013. DOI: 10.1002/9781118481844.

[46] G. E. P. Box and N. R. Draper. *Response Surfaces, Mixtures, and Ridge Analyses*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2007. DOI: 10.1002/0470072768.

[47] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. 4th ed. Wiley Series in Probability and Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2016.

[48] O. P. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Scientific Computation. Dordrecht, Netherlands: Springer, 2010. DOI: 10.1007/978-90-481-3520-2.

[49] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton, New Jersey, USA: Princeton University Press, 2010.

[50] T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. New York: Springer, 2003. DOI: 10.1007/978-1-4757-3799-8.

[51] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts, USA: The MIT Press, 2006.

[52] C. M. Bishop. *Neural Networks for Pattern Recognition*. Advanced Texts in Econometrics. Oxford, UK: Oxford University Press, 1995.

[53] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press, 1996. DOI: 10.1017/CBO9780511812651.

[54] A. Christmann and I. Steinwart. *Support Vector Machines*. Information Science and Statistics. New York: Springer, 2008. DOI: 10.1007/978-0-387-77242-4.

[55] S. Abe. *Support Vector Machines for Pattern Classification*. 2nd ed. Advances in Pattern Recognition. London, UK: Springer-Verlag, 2010. DOI: 10.1007/978-1-84996-098-4.

[56] Y. M. Marzouk, H. N. Najm, and L. A. Rahn. "Stochastic spectral methods for efficient Bayesian solution of inverse problems". In: *Journal of Computational Physics* 224.2 (2007), pp. 560–586. DOI: 10.1016/j.jcp.2006.10.010.

[57] Y. M. Marzouk and H. N. Najm. "Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems". In: *Journal of Computational Physics* 228.6 (2009), pp. 1862–1902. DOI: 10.1016/j.jcp.2008.11.024.

[58] V. R. Joseph. "Bayesian Computation Using Design of Experiments-Based Interpolation Technique". In: *Technometrics* 54.3 (2012), pp. 209–225. DOI: 10.1080/00401706.2012.680399.

[59] V. R. Joseph. "A Note on Nonnegative DoIt Approximation". In: *Technometrics* 55.1 (2013), pp. 103–107. DOI: 10.1080/00401706.2012.759154.

[60] M. Raudenský, J. Horský, J. Krejsa, and L. Sláma. "Usage of Artificial Intelligence Methods in Inverse Problems for Estimation of Material Parameters". In: *International Journal of Numerical Methods for Heat and Fluid Flow* 6.8 (1996), pp. 19–29. DOI: 10.1108/eb017555.

[61] T. Mareš, E. Janouchová, and A. Kučerová. "Artificial neural networks in the calibration of nonlinear mechanical models". In: *Advances in Engineering Software* 95 (2016), pp. 68–81. DOI: 10.1016/j.advengsoft.2016.01.017.

[62] P. Kersaudy, B. Sudret, N. Varsier, O. Picon, and J. Wiart. "A new surrogate modeling technique combining Kriging and polynomial chaos expansions – Application to uncertainty analysis in computational dosimetry". In: *Journal of Computational Physics* 286 (2015), pp. 103–117. DOI: 10.1016/j.jcp.2015.01.034.

[63] R. Schöbi, B. Sudret, and J. Wiart. "Polynomial-chaos-based Kriging". In: *International Journal for Uncertainty Quantification* 5.2 (2015), pp. 171–193. DOI: 10.1615/Int.J.UncertaintyQuantification.2015012467.

[64] M. D. Spiridonakos and E. N. Chatzi. "Metamodeling of dynamic nonlinear structural systems through polynomial chaos NARX models". In: *Computers & Structures* 157 (2015), pp. 99–113. DOI: 10.1016/j.compstruc.2015.05.002.

[65] C. V. Mai, M. D. Spyridonakos, E. N. Chatzi, and B. Sudret. "Surrogate Modeling for Stochastic Dynamical Systems by Combining Nonlinear Autoregressive with Exogenous Input Models and Polynomial Chaos Expansions". In: *International Journal for Uncertainty Quantification* 6.4 (2016), pp. 313–339. DOI: 10.1615/Int.J.UncertaintyQuantification.2016016603.

[66] N. Wiener. "The Homogeneous Chaos". In: *American Journal of Mathematics* 60.4 (1938), pp. 897–936. DOI: 10.2307/2371268.

[67] R. H. Cameron and W. T. Martin. "The Orthogonal Development of Non-Linear Functionals in Series of Fourier-Hermite Functionals". In: *Annals of Mathematics* 48.2 (1947), pp. 385–392. DOI: 10.2307/1969178.

[68] R. Ghanem and P. D. Spanos. "Polynomial Chaos in Stochastic Finite Elements". In: *Journal of Applied Mechanics* 57.1 (1990), pp. 197–202. DOI: 10.1115/1.2888303.

[69] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. New York: Springer-Verlag, 1991. DOI: 10.1007/978-1-4612-3094-6.

[70] T. J. Rivlin. *An Introduction to the Approximation of Functions*. Waltham, Massachusetts, USA: Blaisdell Publishing Company, 1969.

[71] J. de Villiers. *Mathematics of Approximation*. Mathematics Textbooks for Science and Engineering 1. Paris, France: Atlantis Press, 2012. DOI: 10.2991/978-94-91216-50-3.

[72] D. Xiu, D. Lucor, C.-H. Su, and G. E. Karniadakis. "Stochastic Modeling of Flow-Structure Interactions Using Generalized Polynomial Chaos". In: *Journal of Fluids Engineering* 124.1 (2002), pp. 51–59. DOI: 10.1115/1.1436089.

[73] D. Xiu and G. E. Karniadakis. "The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations". In: *SIAM Journal on Scientific Computing* 24.2 (2002), pp. 619–644. DOI: 10.1137/S1064827501387826.

[74] D. Xiu and G. E. Karniadakis. "Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos". In: *Computer Methods in Applied Mechanics and Engineering* 191.43 (2002), pp. 4927–4948. DOI: 10.1016/S0045-7825(02)00421-8.

[75] O. G. Ernst, A. Mugler, H.-J. Starkloff, and E. Ullmann. "On the convergence of generalized polynomial chaos expansions". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 46.2 (2012), pp. 317–339. DOI: 10.1051/m2an/2011045.

[76] J. A. S. Witteveen and H. Bijl. "Modeling Arbitrary Uncertainties Using Gram-Schmidt Polynomial Chaos". In: *44th AIAA Aerospace Sciences Meeting and Exhibit*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2006. DOI: 10.2514/6.2006-896.

[77] S. Oladyshkin and W. Nowak. "Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion". In: *Reliability Engineering & System Safety* 106 (2012), pp. 179–190. DOI: 10.1016/j.ress.2012.05.002.

[78] R. Ahlfeld, B. Belkouchi, and F. Montomoli. "SAMBA: Sparse Approximation of Moment-Based Arbitrary Polynomial Chaos". In: *Journal of Computational Physics* 320 (2016), pp. 1–16. DOI: 10.1016/j.jcp.2016.05.014.

[79] Å. Björck. "The calculation of linear least squares problems". In: *Acta Numerica* 13 (2004), pp. 1–53. DOI: 10.1017/S0962492904000169.

[80] L. Breiman and P. Spector. "Submodel Selection and Evaluation in Regression. The X-Random Case". In: *International Statistical Review* 60.3 (1992), pp. 291–319. DOI: 10.2307/1403680.

[81] S. Borra and A. Di Ciaccio. "Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods". In: *Computational Statistics & Data Analysis* 54.12 (2010), pp. 2976–2989. DOI: 10.1016/j.csda.2010.03.004.

[82] X. Su, X. Yan, and C.-L. Tsai. "Linear regression". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.3 (2012), pp. 275–294. DOI: 10.1002/wics.1198.

[83] V. V. Fedorov and P. Hackl. *Model-Oriented Design of Experiments*. Lecture Notes in Statistics 125. New York: Springer, 1997. DOI: 10.1007/978-1-4612-0703-0.

[84] F. Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics 50. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2006. DOI: 10.1137/1.9780898719109.

[85] N. Ahmed and K. R. Rao. *Orthogonal Transforms for Digital Signal Processing*. Springer-Verlag Berlin Heidelberg, 1975. DOI: 10.1007/978-3-642-45450-9.

[86] R. Wang. *Introduction to Orthogonal Transforms: With Applications in Data Processing and Analysis*. Cambridge, UK: Cambridge University Press, 2012. DOI: 10.1017/CBO9781139015158.

[87] R. Bellman. *Dynamic Programming*. Princeton, New Jersey, USA: Princeton University Press, 1957.

[88] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton, New Jersey, USA: Princeton University Press, 1961.

[89] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer, 2006.

[90] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts, USA: The MIT Press, 2012.

[91] S. Kulkarni and G. Harman. *An Elementary Introduction to Statistical Learning Theory*. Wiley Series in Probability and Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2011. DOI: 10.1002/9781118023471.

[92] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics 103. New York: Springer, 2013. DOI: 10.1007/978-1-4614-7138-7.

[93] A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. New York: Springer, 2008. DOI: 10.1007/978-0-387-78189-1.

[94] I. Koch. *Analysis of Multivariate and High-Dimensional Data*. Cambridge Series in Statistical and Probabilistic Mathematics 32. New York: Cambridge University Press, 2014. DOI: 10.1017/CBO9781139025805.

[95] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. New York: Springer, 2007. DOI: 10.1007/978-0-387-39351-3.

[96] P. G. Constantine, E. Dow, and Q. Wang. "Active Subspace Methods in Theory and Practice: Applications to Kriging Surfaces". In: *SIAM Journal on Scientific Computing*. A 36.4 (2014), pp. 1500–1524. DOI: 10.1137/130916138.

[97] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. New York: Springer, 2013. DOI: 10.1007/978-0-8176-4948-7.

Chapter 3

Bayesian inference

The preceding chapter covered the issue of how parameter uncertainties impact on the model predictions. We now answer the reverse question of how measured output data can inform about the model parameters and reduce their uncertainty. Bayesian inference establishes a probabilistic framework that allows one to coherently quantify uncertainties with due regard to all available information. It is based on the transition of a prior into a posterior probability distribution reflecting the learning process. The prior and the posterior represent the state of knowledge or level of epistemic uncertainty before and after incorporating the experimental data.

Bayesian probability encompasses a whole range of philosophical attitudes, mathematical developments and computational tools for statistical data analysis. Even nowadays it is instructive and occasionally entertaining to have a look into the classical literature [1, 2]. Refreshingly original perspectives are also formulated in [3, 4]. More technical expositions of Bayesian experiments are found in the lesser known textbook [5] or in more contemporary introductions to mathematical statistics [6, 7] and its measure-theoretical foundations [8, 9]. The monograph [10] exclusively addresses conditional distributions that are of particular importance in Bayesian inference. Modern introductions strongly emphasize practical and computational aspects [11, 12].

The popularity of the Bayesian approach is explained by its power to quantify and reduce uncertainties in complex problems. Parameter estimation [13, 14] and data assimilation [15, 16] are a few prototypical tasks. They are important for applications in social [17, 18], physical [19, 20] and engineering sciences [21, 22]. Complex problems, where uncertainty [23–25] and physical modeling [26, 27] take place at multiple system levels, can be solved with Bayesian methods. Interesting examples can be found in Chapters 6, 7 and 10 of the thesis.

For the sake of completeness it is remarked that Bayesian probability is not limited to statistical problems only. Another trendy application is indeed probabilistic numerics [28, 29]. Numerical algorithms can be endowed with a probabilistic interpretation in a way that allows one to quantify the confidence in the computed solutions [30]. In this regard, Bayesian quadrature [31, 32] itself is applicable to Bayesian computations [33, 34].

This introductory chapter on Bayesian inference is structured in the following way. The foundations of statistical modeling are introduced in Sections 3.1 and 3.2, where the likelihood function and the prior distribution are discussed. Inference based on the posterior is subsequently explained in Section 3.3. The question of the model evidence and some of its ramifications are dealt with in Section 3.4. Some practical issues related to the model parametrization are expounded in Section 3.5. Bayesian inverse problems are addressed afterwards in Section 3.6. The numerical computation of the posterior distribution is dissected in Section 3.7.

3.1 Likelihood function

In the following, the unknown parameters of a statistical model are denoted as $\boldsymbol{x} = (x_1, \dots, x_M)^\top \in \mathbb{R}^M$. The number of unknowns is denoted as $M \in \mathbb{N}_{>0}$ which is in line with the notation of the preceding chapter. It is aimed at the statistical identification of the unknown system parameters with $N \in \mathbb{N}_{>0}$ real measurements $\boldsymbol{y} = (y_1, \dots, y_N)^\top \in \mathbb{R}^N$ of related observables. In order to draw inferences from the data $\boldsymbol{y}$ about the unknowns $\boldsymbol{x}$, one has to establish the connection between them. Therefore one constructs a probabilistic model $\pi(\boldsymbol{y}|\boldsymbol{x})$ that explains the randomness of the data for given parameter values. This is often denoted as

$$\boldsymbol{Y}|\boldsymbol{x} \sim \pi(\boldsymbol{y}|\boldsymbol{x}). \tag{3.1}$$

This equation actually may involve fully known covariates, i.e. explanatory control variables or experimental conditions. They are omitted here for the sake of notational convenience. In a way the unknowns $\boldsymbol{x}$ index the data distribution, while the actually acquired data are interpreted as a random realization $\boldsymbol{Y} = \boldsymbol{y}$ generated from Eq. (3.1) for the true values of the unknowns.

It is remarked that while the form of $\pi(\boldsymbol{y}|\boldsymbol{x})$ is seemingly simple, it embodies a variety of assumptions and simplifications that are made during the modeling process. Even the notions of true parameter values and a data generating mechanism can be seen as conceptualizations. Examples of how such statistical data models can be constructed in connection with inverse modeling are discussed in Section 3.6.

The *likelihood function* plays a key role in frequentist and Bayesian inference [35, 36]. For the obtained and therefore fixed observations $\boldsymbol{y}$, it is defined as

$$\mathcal{L}(\boldsymbol{x}) = \pi(\boldsymbol{y}|\boldsymbol{x}). \tag{3.2}$$

Hence, the likelihood emerges from evaluating the conditional density in Eq. (3.1) as a function of the unknowns $\boldsymbol{x}$. In *maximum likelihood estimation* (MLE) the unknown parameters are estimated as

$$\hat{\boldsymbol{x}}_{\mathrm{MLE}} = \underset{\boldsymbol{x}\in\mathbb{R}^M}{\arg\max}\, \mathcal{L}(\boldsymbol{x}). \tag{3.3}$$

The point estimator maximizes the likelihood function in Eq. (3.2) over the admissible parameter values. A weak point of the MLE in Eq. (3.3) is that it does not allow for quantifying the unavoidable statistical uncertainty. This motivates Bayesian inference which is capable of doing so.

3.2 Prior distribution

The Bayesian approach to inference and prediction builds on probabilistic reasoning. This way it allows for a more thorough information processing and uncertainty analysis. Involving a subjective interpretation of probability, randomness is not only attributed to the data $\boldsymbol{y}$ as in Eq. (3.1), but also to the unknowns $\boldsymbol{x}$. The modeler's and analyst's ignorance regarding the true parameter values before analyzing the data is represented as a random vector

$$\boldsymbol{X} \sim \pi(\boldsymbol{x}). \tag{3.4}$$

Here, $\pi(\boldsymbol{x})$ is called the *prior density*. Instead of merely acknowledging the fact that the parameter values are not known, their epistemic uncertainty is modeled as a probability distribution. The true values are then considered a realization $\boldsymbol{X} = \boldsymbol{x}$ of the random vector in Eq. (3.4). Note that the randomness does not refer to draws in a frequentist sense, but to a lack of knowledge in a Bayesian sense. As detailed in Section 3.3, the analyst can update his or her knowledge by conditioning on the realized data. Beforehand it is necessary to construct an appropriate prior distribution.

The selection of the prior distribution is of utmost practical importance in Bayesian inference. It is in fact the most controversial aspect. On the one hand, the prior allows one to incorporate qualitative and quantitative information other than the data. Beyond physical constraints, this includes heterogeneous sources such as expert knowledge, previous experiments and published literature. On the other hand, this raises the question of how to encode such information into a probability distribution. Similar to the assumptions and compromises that have to be made in order to formulate a probabilistic data model as in Eq. (3.1), the determination of the prior parameter model in Eq. (3.4) can be understood as a modeling choice. As such, it may be subject to various guiding principles.

Very generally, one may classify Bayesian priors according to the way they are chosen, the information they convey and the function they fulfill. For a start, one may distinguish between priors that are more *subjective*, i.e. elicited on the basis of one's own or someone else's personal belief [37, 38], or more *objective*, i.e. constructed according to some formal rules [39, 40]. The latter includes the maximum entropy principle [41, 42]. Subjective and objective prior distributions are sometimes also called *informative* and *uninformative*, respectively. More generally, these terms can be used in order to characterize the prior with respect to its information content. Real prior distributions may occupy a wide spectrum that ranges from more subjective or informative to rather objective or uninformative. There are also more or less functional priors that serve certain purposes. They are chosen for mere mathematical convenience or their regularization properties. Conjugacy [43], robustness [44] and sparsity [45–47] can be for instance achieved by choosing appropriate priors.

In engineering practice, one often designates a well-known family of distributions as candidate priors. The corresponding parameters are then set so as to mirror the uncertainty as faithfully as possible. Uniform distributions are often chosen for parameters that can be bounded from above and below, e.g. due to physical constraints. Gaussian or lognormal distributions are often used for parameters that are unbounded or strictly positive, respectively.

3.3 Posterior distribution

All things considered, Bayesian modeling rests on the marginal distribution $\pi(\boldsymbol{x})$ of the unknown parameters in Eq. (3.4) and the conditional distribution $\pi(\boldsymbol{y}|\boldsymbol{x})$ of the observational data in Eq. (3.1). Consequentially the unknowns and the data are represented as jointly distributed random vectors

$$(\boldsymbol{Y}, \boldsymbol{X}) \sim \pi(\boldsymbol{y}, \boldsymbol{x}) = \pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}). \tag{3.5}$$

This is a complete probability model of the Bayesian experiment. The true parameters and the actual data are regarded as a realization $(\boldsymbol{Y}, \boldsymbol{X}) = (\boldsymbol{y}, \boldsymbol{x})$ of the joint random variables in Eq. (3.5). While the outcome of the data $\boldsymbol{y}$ is observed, the true parameters $\boldsymbol{x}$ remain unobserved.

Now one can synthesize the prior information and the observed data in order to estimate the unknowns. In particular, one proceeds by conditioning on the realized data. Given the likelihood function in Eq. (3.2) and the prior density in Eq. (3.4), the *posterior density* follows from Bayes' law

$$\pi(\boldsymbol{x}|\boldsymbol{y}) = \frac{\mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})}{Z}, \tag{3.6}$$

$$Z = \int_{\mathbb{R}^M} \mathcal{L}(\boldsymbol{x})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{3.7}$$

The normalizing constant Z is usually called the *model evidence* or *marginal likelihood*. It will be further examined in Section 3.4. In the same fashion as the prior represents the uncertainty about the unknowns before analyzing the data, the posterior in Eq. (3.6) summarizes the reduced uncertainty afterwards.

In Fig. 3.1 the functioning of Bayesian updating is illustrated for a single quantity of interest (QoI). The prior is transited into the posterior density, which is paralleled by a reduction of the epistemic uncertainty and a higher degree of probability mass localization. For the sake of clarity, both the prior and the posterior density in the sketch are Gaussian. In most but the simplest cases, however, the posterior is a complex probability distribution that may exhibit strong non-normalities and a multiplicity of modes. Multivariate posteriors often contain linear correlations and complex dependencies between the variables involved.

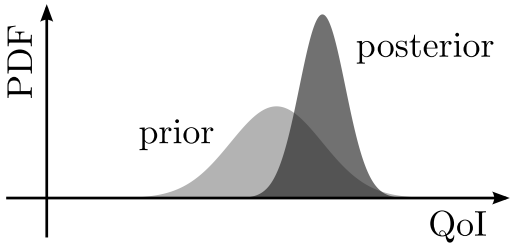


Figure 3.1: Bayesian inference.

Now all information is contained in the posterior probability density function. As opposed to the determination of the prior where the question was how to encode information into a probability density, the question here becomes how to decode it from the posterior. A natural way is to evaluate conditional expectation values and regular probabilities given the data. The expectation of a certain QoI $h\colon \mathbb{R}^M \to \mathbb{R}$ under the posterior can be written as

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \int_{\mathbb{R}^M} h(\boldsymbol{x})\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}. \tag{3.8}$$

Most relevant quantities follow thereby from considering appropriate QoIs. For instance, with the indicator function $h = I_B$ of a set $B \in \mathcal{B}(\mathbb{R}^M)$ one obtains the posterior probability of the event $\boldsymbol{X} \in B$ as

$$\mathbb{P}_{\boldsymbol{X}|\boldsymbol{Y}}(B|\boldsymbol{y}) = \int_B \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}. \tag{3.9}$$

This completes the characterization of the posterior probability distribution. Possibilities to further analyze and summarize the posterior are discussed below.

3.3.1 Posterior summaries

Very often one is interested in the first statistical moments of the posterior. They serve as brief summaries of the possibly complex probability distribution. The expected value and covariance matrix are given as

$$\mathbb{E}[\boldsymbol{X}|\boldsymbol{y}] = \int_{\mathbb{R}^M} \boldsymbol{x}\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}, \tag{3.10}$$

$$\mathrm{Cov}[\boldsymbol{X}|\boldsymbol{y}] = \int_{\mathbb{R}^M} (\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{y}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{y}])^\top\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}. \tag{3.11}$$

The posterior mean in Eq. (3.10) is often taken as a point estimate of the unknown parameter vector, whereas the covariance in Eq. (3.11) is regarded as a measure of the statistical uncertainty. An alternative for point estimation is the *maximum a posteriori* (MAP) estimate. It maximizes the posterior density over the admissible parameter values. Simply put, this is just the mode

$$\hat{\boldsymbol{x}}_{\mathrm{MAP}} = \underset{\boldsymbol{x}\in\mathbb{R}^M}{\arg\max}\, \pi(\boldsymbol{x}|\boldsymbol{y}). \tag{3.12}$$

For interval estimation one usually specifies *credible regions* that accumulate a certain high percentage of the total posterior mass. The probability that such sets indeed contain the true parameter values is determined according to Eq. (3.9).

Bayesian point estimation can be more formally understood within a decision-theoretic framework [48, 49]. A *Bayes estimator* minimizes/maximizes the expected value of a certain loss/utility function under the posterior distribution. The posterior mean in Eq. (3.10) is the Bayes estimator for a quadratic risk function. Technically speaking, MAP estimation in Eq. (3.12) does not establish a proper Bayes estimator. It can be understood as a limit of such, though.

In the multivariate case, the joint posterior may contain dependencies between the components of $\boldsymbol{x}$. Although the presence or lack of such structures gives deep insight into the problem at hand, one may want to disregard them for the moment and focus on the posterior marginals. For a specific parameter x_i with $i \in \{1, \dots, M\}$ one integrates the posterior over the remaining unknowns $\boldsymbol{x}_{\sim i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_M)^\top$ in order to obtain

$$\pi(x_i|\boldsymbol{y}) = \int_{\mathbb{R}^{M-1}} \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}_{\sim i}. \tag{3.13}$$

This summarizes the accumulated information on the parameter x_i. Since one-dimensional posterior marginals can be visualized nicely, they are often plotted as graphical summaries of the joint posterior.

3.3.2 Predictive distributions

In the same way as the prior $\pi(\boldsymbol{x})$ informs about $\boldsymbol{x}$, i.e. the distribution represents a subjective uncertainty rather than a sampling frequency, our expectations on the data $\boldsymbol{y}$ before seeing them are summarized by

$$\pi(\boldsymbol{y}) = \int_{\mathbb{R}^M} \pi(\boldsymbol{y}|\boldsymbol{x})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{3.14}$$

This is called the *prior predictive density*. The model evidence in Eq. (3.7) actually stems from evaluating the prior predictive density at the real data. After analyzing the data $\boldsymbol{y}$, one can similarly predict the future outcome $\boldsymbol{y}'$ in a replication of the experiment.

It is often assumed that the observed and the future data are conditionally independent. This means that $\pi(\boldsymbol{y}, \boldsymbol{y}'|\boldsymbol{x}) = \pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{y}'|\boldsymbol{x})$. The probabilistic data model in Eq. (3.1) can then be adjusted to as yet unobserved data $\boldsymbol{Y}' \sim \pi(\boldsymbol{y}'|\boldsymbol{x})$, e.g. by resetting the covariates. One could simply use $\pi(\boldsymbol{y}'|\hat{\boldsymbol{x}})$ for predicting the future data, where $\hat{\boldsymbol{x}}$ is some point estimate of the parameters that comes from an analysis of the old data. This, however, ignores the statistical estimation uncertainty. It is therefore advisable to average over the posterior so as to derive the *posterior predictive density*

$$\pi(\boldsymbol{y}'|\boldsymbol{y}) = \int_{\mathbb{R}^M} \pi(\boldsymbol{y}'|\boldsymbol{x}, \boldsymbol{y})\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^M} \pi(\boldsymbol{y}'|\boldsymbol{x})\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}. \tag{3.15}$$

The second equality is a direct consequence of the conditional independence $\pi(\boldsymbol{y}'|\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}'|\boldsymbol{x})$. Similar to the unconditional data predictions in Eq. (3.14), the distribution in Eq. (3.15) predicts new data given the already observed ones.

3.3.3 Information gain

It is interesting to look at the Bayesian update from an information-theoretic point of view [50, 51]. This is of paramount importance for Bayesian optimal design [52, 53] and the definition of least-informative reference priors [54, 55]. A more specific perspective on the pervasive concepts of "information" and "uncertainty" and their embodiment in probability distributions is entailed hereby [56, 57].

Intuitively one may think of information as the complement of uncertainty. More formally it can be interpreted as the reduction of uncertainty $-\log \pi(\boldsymbol{x})$ that results from receiving the outcome $\boldsymbol{X} = \boldsymbol{x}$ of a random variable $\boldsymbol{X} \sim \pi(\boldsymbol{x})$. Note that this does not describe a Bayesian learning procedure. The *Shannon entropy* quantifies the potential information gain or average surprisal $\mathbb{E}[-\log \pi(\boldsymbol{X})]$ of this communication process [58, 59]. This way it rigorously measures the degree of unpredictability or uncertainty that is inherent in a whole probability distribution. The entropy of the continuous prior density $\pi(\boldsymbol{x})$ is defined as

$$H_{\mathrm{S}}(\pi(\cdot)) = -\int_{\mathbb{R}^M} \log(\pi(\boldsymbol{x}))\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{3.16}$$

Likewise one can define the information entropy $H_{\mathrm{S}}(\pi(\cdot|\boldsymbol{y}))$ of the posterior density $\pi(\boldsymbol{x}|\boldsymbol{y})$. Notice that the differential entropy in Eq. (3.16) may become negative and is not invariant under a change of variables. Actually, these undesirable properties have arisen due to the transfer of the original definition from discrete to continuous random variables.

A relative entropy concept that works equally well in the discrete and the continuous case is the *Kullback–Leibler* (KL) *divergence* [60, 61]. It measures the additional entropy that is introduced when using an approximate or distorted distribution in place of the true reference one. While the actual events follow the reference, they are incorrectly expected according to the approximation. The additional uncertainty of predicting the posterior $\pi(\boldsymbol{x}|\boldsymbol{y})$ as the reference with the prior $\pi(\boldsymbol{x})$ as the approximation can be defined as

$$D_{\mathrm{KL}}(\pi(\cdot|\boldsymbol{y})\|\pi(\cdot)) = \int_{\mathbb{R}^M} \log\left(\frac{\pi(\boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x})}\right) \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} = -\int_{\mathbb{R}^M} \log(\pi(\boldsymbol{x}))\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} - H_{\mathrm{S}}(\pi(\cdot|\boldsymbol{y})). \tag{3.17}$$

The cross-entropy $H_{\mathrm{C}}(\pi(\cdot|\boldsymbol{y}), \pi(\cdot)) = -\int_{\mathbb{R}^M} \log(\pi(\boldsymbol{x}))\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}$ measures the total uncertainty of using the prior instead of the true posterior. Hence, the relative entropy in Eq. (3.17) is the additional uncertainty $D_{\mathrm{KL}}(\pi(\cdot|\boldsymbol{y})\|\pi(\cdot)) = H_{\mathrm{C}}(\pi(\cdot|\boldsymbol{y}), \pi(\cdot)) - H_{\mathrm{S}}(\pi(\cdot|\boldsymbol{y}))$. It is a non-negative and transformation-invariant measure of the difference between two probability distributions in terms of their entropy contents. Moreover, the divergence is asymmetric by construction and attains zero if and only if $\pi(\boldsymbol{x}|\boldsymbol{y}) = \pi(\boldsymbol{x})$.

Following these remarks, one can interpret the KL divergence $D_{\mathrm{KL}}(\pi(\cdot|\boldsymbol{y})\|\pi(\cdot))$ as a measure of the uncertainty reduction that comes along with the passage from the prior to the posterior. This may be seen as the amount of information brought by the data in turn. It is worth mentioning here that this measure of the information gain is never negative, no matter of how the distributions look like, and regardless of whether the divergence from the prior to the posterior or from the posterior to the prior is considered.

3.4 Model evidence

In this section we briefly address the basics of multi-model inference [62–64]. This subsumes Bayesian model comparison, selection and averaging that are important when there is an abundance of models available for predicting the data [65–67]. Although these issues are not the main topics of this dissertation, they are revelatory about the probabilistic rationale of single-model parameter calibration.

Sometimes a whole set of $L \in \mathbb{N}_{>1}$ candidate models $\boldsymbol{\mathcal{H}} = \{\mathcal{H}_1, \dots, \mathcal{H}_L\}$ with densities $\pi(\boldsymbol{y}|\boldsymbol{x}_{\mathcal{H}}, \mathcal{H})$ as in Eq. (3.1) possibly explain the data. As in Eq. (3.4), each model $\mathcal{H} \in \boldsymbol{\mathcal{H}}$ has its own tuning parameters $\boldsymbol{x}_{\mathcal{H}} \in \mathbb{R}^{M_{\mathcal{H}}}$ with $M_{\mathcal{H}} \in \mathbb{N}_{>0}$ whose prior uncertainty is represented by $\pi(\boldsymbol{x}_{\mathcal{H}}|\mathcal{H})$. Consequentially, prior predictive distributions of the form as in Eq. (3.14) can be constructed as

$$\pi(\boldsymbol{y}|\mathcal{H}) = \int_{\mathbb{R}^{M_{\mathcal{H}}}} \pi(\boldsymbol{y}|\boldsymbol{x}_{\mathcal{H}}, \mathcal{H})\, \pi(\boldsymbol{x}_{\mathcal{H}}|\mathcal{H})\, \mathrm{d}\boldsymbol{x}_{\mathcal{H}}. \tag{3.18}$$

The actual data $\boldsymbol{y}$ can be used in order to estimate the unknowns $\boldsymbol{x}_{\mathcal{H}}$ for each model separately. Posteriors $\pi(\boldsymbol{x}_{\mathcal{H}}|\boldsymbol{y}, \mathcal{H})$ as in Eq. (3.6) arise in that context. One can advance probabilistic reasoning to the next higher level by considering *model uncertainty*. A discrete probability distribution $\pi(\mathcal{H})$ is assigned over the candidate models $\mathcal{H} \in \boldsymbol{\mathcal{H}}$. This distribution characterizes the prior plausibility of the hypothesis that $\mathcal{H}$ is the true model.

Conditioning on the collected data $\boldsymbol{y}$ results in the corresponding posterior probabilities of the hypothesized models

$$\pi(\mathcal{H}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\mathcal{H})\pi(\mathcal{H})}{\sum_{\mathcal{H}'\in\boldsymbol{\mathcal{H}}}\pi(\boldsymbol{y}|\mathcal{H}')\,\pi(\mathcal{H}')}. \tag{3.19}$$

The role of the likelihood function is filled by the conditional density in Eq. (3.18) which, for fixed data, is evaluated as a function of the model. According to Eq. (3.7) this is exactly the $\mathcal{H}$-specific evidence $Z_{\mathcal{H}} = \pi(\boldsymbol{y}|\mathcal{H})$.

Beyond revealing the significance of the single-model posterior normalization constant for multi-model inference, this also highlights the relative character of Bayesian probabilities in general. They are conditional on the modeling assumptions made and have to be assessed with respect to the alternative hypotheses tested. In the single-model case, the continuous posterior in Eq. (3.6) weighs possible parameter values given one specific statistical model. In the multi-model case, the discrete posterior in Eq. (3.19) compares statistical models against a number of predefined candidates.

Model comparison and *model selection* can be based on Eqs. (3.18) and (3.19). The posterior mode $\hat{\mathcal{H}} = \arg\max_{\mathcal{H}\in\boldsymbol{\mathcal{H}}}\pi(\mathcal{H}|\boldsymbol{y})$ is the best model as suggested by the data and the prior information. *Bayes factors* are often used for hypothesis testing and pairwise model comparison [68, 69]. For two models, say $\mathcal{H}_1$ and $\mathcal{H}_2$, they are defined as the marginal likelihood odds

$$B_{1,2} = \frac{\pi(\boldsymbol{y}|\mathcal{H}_1)}{\pi(\boldsymbol{y}|\mathcal{H}_2)} = \frac{\int_{\mathbb{R}^{M_{\mathcal{H}_1}}}\pi(\boldsymbol{y}|\boldsymbol{x}_{\mathcal{H}_1},\mathcal{H}_1)\,\pi(\boldsymbol{x}_{\mathcal{H}_1}|\mathcal{H}_1)\,\mathrm{d}\boldsymbol{x}_{\mathcal{H}_1}}{\int_{\mathbb{R}^{M_{\mathcal{H}_2}}}\pi(\boldsymbol{y}|\boldsymbol{x}_{\mathcal{H}_2},\mathcal{H}_2)\,\pi(\boldsymbol{x}_{\mathcal{H}_2}|\mathcal{H}_2)\,\mathrm{d}\boldsymbol{x}_{\mathcal{H}_2}}. \tag{3.20}$$

This measures the evidence of $\mathcal{H}_1$ with respect to the alternative model $\mathcal{H}_2$. In case of a uniform prior with $\pi(\mathcal{H}_1) = \pi(\mathcal{H}_2)$, the Bayes factor in Eq. (3.20) equals the posterior odds $B_{1,2} = \pi(\mathcal{H}_1|\boldsymbol{y})/\pi(\mathcal{H}_2|\boldsymbol{y})$.

Bayesian model selection provides an automatic implementation of *Occam's razor*, i.e. the common sense that simple models should be preferred over complex ones if they equally well explain the data. This is sometimes interpreted as a formal justification of parsimony as a principle. Note that simplicity here does not primarily refer to the intricacy of the mathematics or solely to the number of the parameters involved. Rather it relates to the predictive spread of the models over their parametric prior uncertainty. The prior predictive distributions in Eq. (3.18) have to integrate to one $\int_{\mathbb{R}^N}\pi(\boldsymbol{y}|\mathcal{H})\,\mathrm{d}\boldsymbol{y} = 1$. Hence, models $\mathcal{H}$ for which the predictions according to $\pi(\boldsymbol{y}|\mathcal{H})$ occupy larger proportions of the data space $\mathbb{R}^N$ have a tendency to lower evidences [70, 71]. In this sense, complex models are naturally penalized on the evidence level $Z_{\mathcal{H}}$. This obviates the need for an ad hoc discrimination on the prior level $\pi(\mathcal{H})$.

A more quantitative understanding of Occam's razor in the context of Bayesian model selection is obtained as follows. Bayes' rule $\pi(\boldsymbol{x}_{\mathcal{H}}|\boldsymbol{y},\mathcal{H}) = Z_{\mathcal{H}}^{-1}\pi(\boldsymbol{y}|\boldsymbol{x}_{\mathcal{H}},\mathcal{H})\pi(\boldsymbol{x}_{\mathcal{H}}|\mathcal{H})$ can be plugged into the definition of the KL divergence $D_{\mathrm{KL}}(\pi(\cdot|\boldsymbol{y},\mathcal{H})\|\pi(\cdot|\mathcal{H}))$ between the posterior and the prior in Eq. (3.17). By solving for the log-evidence one easily derives

$$\log Z_{\mathcal{H}} = \int_{\mathbb{R}^{M_{\mathcal{H}}}}\log(\pi(\boldsymbol{y}|\boldsymbol{x}_{\mathcal{H}},\mathcal{H}))\,\pi(\boldsymbol{x}_{\mathcal{H}}|\boldsymbol{y},\mathcal{H})\,\mathrm{d}\boldsymbol{x}_{\mathcal{H}} - D_{\mathrm{KL}}(\pi(\cdot|\boldsymbol{y},\mathcal{H})\|\pi(\cdot|\mathcal{H})). \tag{3.21}$$

The first term in Eq. (3.21) is the posterior expectation of the log-likelihood and therefore favors models that fit the data well [72, 73]. The second term is the relative entropy that penalizes complex models for which the uncertainty reduction is high [74, 75].

With a model-specific posterior distribution of the unknown parameters $\pi(\boldsymbol{x}_{\mathcal{H}}|\boldsymbol{y},\mathcal{H})$, one can base predictions of future data on an adapted model as in Eq. (3.15). However, in the multi-model context one has a number of $\mathcal{H}$-specific predictive distributions

$$\pi(\boldsymbol{y}'|\boldsymbol{y},\mathcal{H}) = \int_{\mathbb{R}^{M_{\mathcal{H}}}}\pi(\boldsymbol{y}'|\boldsymbol{x}_{\mathcal{H}},\mathcal{H})\,\pi(\boldsymbol{x}_{\mathcal{H}}|\boldsymbol{y},\mathcal{H})\,\mathrm{d}\boldsymbol{x}_{\mathcal{H}}. \tag{3.22}$$

One way to predict new data is to simply use the predictive distribution $\pi(\boldsymbol{y}'|\boldsymbol{y},\hat{\mathcal{H}})$ of the MAP model $\hat{\mathcal{H}}$. Another more coherent way would be *Bayesian model averaging* [76, 77], where the distribution of future data is expressed as the posterior-weighted average

$$\pi(\boldsymbol{y}'|\boldsymbol{y}) = \sum_{\mathcal{H}\in\boldsymbol{\mathcal{H}}}\pi(\boldsymbol{y}'|\boldsymbol{y},\mathcal{H})\,\pi(\mathcal{H}|\boldsymbol{y}). \tag{3.23}$$

The distribution in Eq. (3.23) is actually a mixture of the components $\pi(\boldsymbol{y}'|\boldsymbol{y},\mathcal{H})$ in Eq. (3.22). This fully acknowledges both the parameter estimation and the model selection uncertainty.

3.5 Model parametrization

The parametrization of Bayesian models and their reparametrization are issues of high practical relevance. Sometimes a model features a more or less natural parametrization, which our notation actually suggests so far. In other cases it may be dictated by the goal of the analysis. Very often, however, there are various equivalent ways of parametrizing the problem and it may not be clear which one is favorable. In any case, it may be convenient to solve a problem not directly but after a suitably chosen transformation. This could either apply to the parameters [78] or to the data [79, 80]. Since such issues are often left aside, in this section we shortly discuss parameter transformations and some related invariance principles.

We will come across a parametrization issue in Section 3.6.2, where the residual model could be either parametrized by the variance or a standard deviation. Moreover, parameter transformations will be important for Chapter 8 in the context of standardized prior distributions and their associated orthogonal polynomials. Some of the material will be also relevant for Chapter 9 in the context of finding a transformation between two given density functions.

Consider a model reparametrization based on a sufficiently well-behaved one-to-one parameter transformation $\mathcal{T}\colon \mathbb{R}^M \to \mathbb{R}^M$ with $\tilde{\boldsymbol{x}} = \mathcal{T}(\boldsymbol{x})$ and $\boldsymbol{x} = \mathcal{T}^{-1}(\tilde{\boldsymbol{x}})$. If $\boldsymbol{X} \sim \pi(\boldsymbol{x})$ is distributed according to the prior, the density of the transformed random variable $\tilde{\boldsymbol{X}} = \mathcal{T}(\boldsymbol{X}) \sim \pi_{\mathcal{T}}(\tilde{\boldsymbol{x}})$ can be written as the *change of variables*

$$\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}}) = \pi(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))\,|\det \boldsymbol{J}_{\mathcal{T}^{-1}}(\tilde{\boldsymbol{x}})|. \tag{3.24}$$

Here, the *Jacobian matrix* is denoted as $\boldsymbol{J}_{\mathcal{T}^{-1}} = \mathrm{d}\mathcal{T}^{-1}/\mathrm{d}\tilde{\boldsymbol{x}}$. The transformation of the prior density in Eq. (3.24) is defined such that one can write a general prior expectation $\mathbb{E}[h(\boldsymbol{X})]$ as the *integration by substitution*

$$\mathbb{E}[h(\boldsymbol{X})] = \int_{\mathbb{R}^M} h(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^M} h(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))\,\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}})\,\mathrm{d}\tilde{\boldsymbol{x}}. \tag{3.25}$$

The transformation of the posterior density can be obtained in a condition-then-transform or transform-then-condition fashion. Similar to Eq. (3.24), one can write this density as

$$\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}}|\boldsymbol{y}) = \pi(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}})|\boldsymbol{y})\,|\det \boldsymbol{J}_{\mathcal{T}^{-1}}(\tilde{\boldsymbol{x}})| = \frac{\mathcal{L}(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))\pi(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))}{Z}\,|\det \boldsymbol{J}_{\mathcal{T}^{-1}}(\tilde{\boldsymbol{x}})| = \frac{\mathcal{L}(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}})}{Z}. \tag{3.26}$$

Analogous to Eq. (3.25), the transformation in Eq. (3.26) is such that one can write a general posterior expectation $\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}]$ as either of the integrals

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \int_{\mathbb{R}^M} h(\boldsymbol{x})\,\pi(\boldsymbol{x}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^M} h(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))\,\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}}|\boldsymbol{y})\,\mathrm{d}\tilde{\boldsymbol{x}}. \tag{3.27}$$

The principles of Bayesian model reparametrization are now fully established. While Eqs. (3.24) and (3.26) represent the transformation of the instrumental densities, the invariance of the relevant expectations is warranted by Eqs. (3.25) and (3.27). In practice, after performing the analysis for the transformed variables, one can back-transform accordingly. However, given certain rules of how to construct prior distributions or how to report point estimates, the results are not generally invariant under parameter transformations.

3.5.1 Invariant priors

Occasionally one may want to define an uninformative prior that reflects a general state of ignorance regarding the true parameter values, i.e. not any admissible value $\boldsymbol{x}$ is favored over others. A typical example is a uniform prior distribution over a bounded support that only discriminates between values inside and outside of the bounds. If $\pi(\boldsymbol{x})$ is such an uninformative prior, then it is interesting to note that after a transformation in Eq. (3.24) the prior $\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}})$ may be actually informative, i.e. it gives undue preference to certain values of $\tilde{\boldsymbol{x}}$. In this sense, if the same recipe of constructing uninformative priors is applied to the parameters $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}} = \mathcal{T}(\boldsymbol{x})$ separately, then one obtains genuinely different Bayesian models. The priors $\pi(\boldsymbol{x})$ and $\tilde{\pi}(\tilde{\boldsymbol{x}})$ chosen that way and the respective posteriors $\pi(\boldsymbol{x}|\boldsymbol{y}) = Z^{-1}\mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})$ and $\tilde{\pi}(\tilde{\boldsymbol{x}}|\boldsymbol{y}) = \tilde{Z}^{-1}\mathcal{L}(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))\tilde{\pi}(\tilde{\boldsymbol{x}})$ are not just different versions of one another modulo the parameter transformation $\mathcal{T}$.

However, one may select priors based on certain invariance principles [81], e.g. *Jeffreys prior* [82] is based on reparametrization invariance. The form of the prior remains unaltered under a change of variables. If $\pi(\boldsymbol{x})$ and $\tilde{\pi}(\tilde{\boldsymbol{x}})$ are independently chosen as this invariant prior, then one has $\tilde{\pi}(\tilde{\boldsymbol{x}}) = \pi_{\mathcal{T}}(\tilde{\boldsymbol{x}}) = \pi(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))\,|\det \boldsymbol{J}_{\mathcal{T}^{-1}}(\tilde{\boldsymbol{x}})|$ for the transformation $\mathcal{T}$ in Eq. (3.24). As a result of this, the posterior $\pi(\boldsymbol{x}|\boldsymbol{y}) = Z^{-1}\mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})$ is the same as $\tilde{\pi}(\tilde{\boldsymbol{x}}|\boldsymbol{y}) = Z^{-1}\mathcal{L}(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}}))\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}})$ up to the change of variables $\tilde{\pi}(\tilde{\boldsymbol{x}}|\boldsymbol{y}) = \pi_{\mathcal{T}}(\tilde{\boldsymbol{x}}|\boldsymbol{y})$ in Eq. (3.26).

3.5.2 Invariant estimators

One can investigate the behavior of point estimators under parameter transformations. If the same principle of estimating $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ is independently applied before and after the transformation, it is natural to ask for the relationship between $\hat{\boldsymbol{x}}$ and $\hat{\tilde{\boldsymbol{x}}}$. An estimation procedure is called *reparametrization invariant* if one has that $\hat{\boldsymbol{x}} = \mathcal{T}^{-1}(\hat{\tilde{\boldsymbol{x}}})$. In a non-Bayesian context, this is a property of the maximum likelihood estimates

$$\hat{\boldsymbol{x}}_{\mathrm{MLE}} = \operatorname*{arg\,max}_{\boldsymbol{x}\in\mathbb{R}^M} \mathcal{L}(\boldsymbol{x}), \quad \hat{\tilde{\boldsymbol{x}}}_{\mathrm{MLE}} = \operatorname*{arg\,max}_{\tilde{\boldsymbol{x}}\in\mathbb{R}^M} \mathcal{L}(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}})). \tag{3.28}$$

In a Bayesian context, applying the same estimation principle to $\pi(\boldsymbol{x}|\boldsymbol{y})$ and $\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}}|\boldsymbol{y})$ generally leads to different point estimates, even if the parameter transformation is accounted for. Due to the Jacobian determinant $\det \boldsymbol{J}_{\mathcal{T}^{-1}}(\tilde{\boldsymbol{x}})$, unlike as in Eq. (3.28), the posterior mode is not transformation invariant. One has $\hat{\boldsymbol{x}}_{\mathrm{MAP}} \neq \mathcal{T}^{-1}(\hat{\tilde{\boldsymbol{x}}}_{\mathrm{MAP}})$ for

$$\hat{\boldsymbol{x}}_{\mathrm{MAP}} = \operatorname*{arg\,max}_{\boldsymbol{x}\in\mathbb{R}^M} \pi(\boldsymbol{x}|\boldsymbol{y}), \quad \hat{\tilde{\boldsymbol{x}}}_{\mathrm{MAP}} = \operatorname*{arg\,max}_{\tilde{\boldsymbol{x}}\in\mathbb{R}^M} \pi(\mathcal{T}^{-1}(\tilde{\boldsymbol{x}})|\boldsymbol{y})\,|\det \boldsymbol{J}_{\mathcal{T}^{-1}}(\tilde{\boldsymbol{x}})|. \tag{3.29}$$

Since the unconditional expectation operator as well as the conditional expectation do not commute with a nonlinear transformation $\mathbb{E}[\mathcal{T}(\boldsymbol{X})|\boldsymbol{y}] \neq \mathcal{T}(\mathbb{E}[\boldsymbol{X}|\boldsymbol{y}])$, neither the posterior mean establishes an invariant estimator. This means $\hat{\boldsymbol{x}} \neq \mathcal{T}^{-1}(\hat{\tilde{\boldsymbol{x}}})$ with

$$\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{y}], \quad \hat{\tilde{\boldsymbol{x}}} = \mathbb{E}[\tilde{\boldsymbol{X}}|\boldsymbol{y}] = \mathbb{E}[\mathcal{T}(\boldsymbol{X})|\boldsymbol{y}]. \tag{3.30}$$

Note that these transformation behaviors are not properties of the posterior, but of the Bayesian point estimates in Eqs. (3.29) and (3.30). One could make analogous statements when applying these estimation principles, i.e. maximizing the density and taking the expected value, to the priors $\pi(\boldsymbol{x})$ and $\pi_{\mathcal{T}}(\tilde{\boldsymbol{x}})$ as well.

3.6 Inverse problems

Inverse modeling is discussed next. This challenging class of problems is important in many areas of science and engineering. Comprehensive overviews on inverse problems from a traditionally deterministic perspective are found in [83, 84]. Inverse problems can be also looked at from a more statistical and Bayesian viewpoint [85, 86]. Notable fields in which classical inverse problems are studied include geophysics [87, 88] as well as earth sciences in general [89, 90], Moreover, inverse problems are important in imaging science [91, 92], scattering theory [93, 94], heat transfer [95, 96] and engineering mechanics [97–99]. Even though the terminology generally differs, a classical inverse problem in engineering sciences is finite element updating [100–102]. At the moment, Bayesian inverse problems receive considerable attention also from the applied mathematics community [103, 104].

An *inverse problem* is posed whenever quantities that cannot be observed directly are determined based on measurements of related quantities. The quantities that interest focuses on and the ones that can be observed are only indirectly connected through a deterministic model. A problem is called *well-posed* after Hadamard if existence, uniqueness and stability of a solution are given. Physical forward problems are often well-posed in this sense. However, inverse problems are typically *ill-posed*, i.e. a solution may be neither existent nor unique, moreover, it may not be continuously dependent on the data. Therefore, such problems have to be *regularized* [105–107]. Treating an inverse problem in a Bayesian frame and imposing a prior distribution might be viewed as a certain regularization procedure [108].

The function $\mathcal{M}\colon \mathbb{R}^M \times \mathbb{R}^D \to \mathbb{R}^N$ that relates the variables $\boldsymbol{x} \in \mathbb{R}^M$ and $\boldsymbol{d} \in \mathbb{R}^D$ with $M, D \in \mathbb{N}_{>0}$ to the observables $\tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{x}) \in \mathbb{R}^N$ is called the *forward model*. It is often nonlinearly dependent on the input parameters. Here, $\boldsymbol{x}$ represents the unknowns as before, while $\boldsymbol{d}$ represents the controllable or at least well-known experimental conditions. Sometimes they are absorbed into the definition of the function $\mathcal{M}$, however, we denote them here explicitly for the sake of clarity. Actual measurements $\boldsymbol{y}$ of the model outputs are then interpreted as the sum

$$\boldsymbol{y} = \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}) + \boldsymbol{\varepsilon} \tag{3.31}$$

of the forward model response $\mathcal{M}(\boldsymbol{x})$ at the true parameter values and a *residual* $\boldsymbol{\varepsilon} \in \mathbb{R}^N$. The latter represents measurement noise and prediction errors. In *statistical inversion*, the residual term is modeled as a random vector $\boldsymbol{E}$. In order to explain the observed data in Eq. (3.31), one thinks of a specific realization $\boldsymbol{E} = \boldsymbol{\varepsilon}$. A common residual model is based on a Gaussian distribution

$$\boldsymbol{E} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{3.32}$$

Neglecting a systematic bias for the moment, here the residuals are centered around $\mathbf{0}$. The symmetric and positive-definite covariance matrix $\boldsymbol{\Sigma}$ characterizes the random errors. It may depend in some form $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{d})$ on the experimental conditions. With the Gaussian residual in Eq. (3.32), the data in Eq. (3.31) is actually seen as realization $\boldsymbol{Y} = \boldsymbol{y}$ of the random variable

$$\boldsymbol{Y}\,|\,\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}|\mathcal{M}(\boldsymbol{x},\boldsymbol{d}),\boldsymbol{\Sigma}). \tag{3.33}$$

Thus, the data are Gaussianly distributed around the model prediction $\mathcal{M}(\boldsymbol{x})$ at the true parameter values. Corresponding to the data model in Eq. (3.33), for the likelihood function in Eq. (3.2) one finds

$$\mathcal{L}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \mathcal{M}(\boldsymbol{x},\boldsymbol{d})\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \mathcal{M}(\boldsymbol{x},\boldsymbol{d})\right)\right). \tag{3.34}$$

Maximizing the likelihood in Eq. (3.34) as in Eq. (3.3) leads to a point estimate $\hat{\boldsymbol{x}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{x}\in\mathbb{R}^M} \mathcal{L}(\boldsymbol{x})$ only. In *Bayesian inversion* one quantifies the statistical estimation uncertainty more thoroughly. The true values of the unknowns are thought of as a realization $\boldsymbol{X} = \boldsymbol{x}$ of a random vector $\boldsymbol{X} \sim \pi(\boldsymbol{x})$ as in Eq. (3.4). A priori, the marginal distribution $\pi(\boldsymbol{x})$ gathers the information about the true parameters. The data are regarded as a realization $\boldsymbol{Y} = \boldsymbol{y}$ of the random vector

$$\boldsymbol{Y} = \mathcal{M}(\boldsymbol{X},\boldsymbol{d}) + \boldsymbol{E}. \tag{3.35}$$

Here, the residual $\boldsymbol{E}$ in Eq. (3.35) is commonly assumed to be statistically independent from the unknowns $\boldsymbol{X}$. With the likelihood $\mathcal{L}(\boldsymbol{x})$ and the prior $\pi(\boldsymbol{x})$, the posterior density $\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})$ as in Eq. (3.6) quantifies the uncertainty of the unknown parameters a posteriori.

The principle of Bayesian inverse problems is illustrated in Fig. 3.2. Contrary to the forward uncertainty quantification problem discussed in Section 2.1 and visualized in Fig. 2.3, the epistemic uncertainty of the unknown but constant parameters is not propagated through the model, but reduced by analyzing a limited number of noisy output measurements. A more exhaustive framework for inverse uncertainty quantification is presented in Chapters 4 and 5, where the variability in the data is backpropagated through the model so as to infer the distribution of aleatory input variables that vary throughout the experiment.

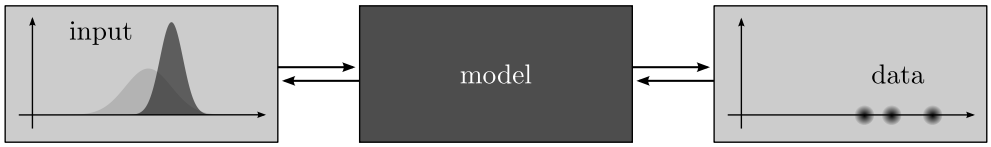


Figure 3.2: Inverse uncertainty quantification.

After having established the posterior distribution of the unknown parameters one is typically interested in making predictions. In a straightforward way, a point estimate $\hat{\boldsymbol{x}}$ may be used in order to predict the system response $\tilde{\boldsymbol{y}}' = \mathcal{M}(\hat{\boldsymbol{x}},\boldsymbol{d}')$ under untested conditions $\boldsymbol{d}' \in \mathbb{R}^D$. Analogous to Eqs. (2.9) and (2.10), one may be also interested in the propagated posterior uncertainty, e.g. the mean response and the covariance matrix

$$\mathbb{E}[\mathcal{M}(\boldsymbol{X},\boldsymbol{d}')|\boldsymbol{y}] = \int_{\mathbb{R}^M} \mathcal{M}(\boldsymbol{x},\boldsymbol{d}')\,\pi(\boldsymbol{x}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}, \tag{3.36}$$

$$\mathrm{Cov}[\mathcal{M}(\boldsymbol{X},\boldsymbol{d}')|\boldsymbol{y}] = \int_{\mathbb{R}^M} \left(\mathcal{M}(\boldsymbol{x},\boldsymbol{d}') - \mathbb{E}[\mathcal{M}(\boldsymbol{X},\boldsymbol{d}')|\boldsymbol{y}]\right)^2 \pi(\boldsymbol{x}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}. \tag{3.37}$$

While Eqs. (3.36) and (3.37) take account of the remaining parameter uncertainty, they leave the output and measurement errors aside. Those are incorporated into the posterior predictive distribution

$$\pi(\boldsymbol{y}'|\boldsymbol{y}) = \int_{\mathbb{R}^M} \mathcal{N}(\boldsymbol{y}'|\mathcal{M}(\boldsymbol{x},\boldsymbol{d}'),\boldsymbol{\Sigma}')\,\pi(\boldsymbol{x}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}. \tag{3.38}$$

New data $\boldsymbol{y}'$ can be predicted for various conditions $\boldsymbol{d}'$ and residual covariances $\boldsymbol{\Sigma}'$ this way. Note that the independence of the errors $\pi(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}') = \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0},\boldsymbol{\Sigma})\,\mathcal{N}(\boldsymbol{\varepsilon}'|\mathbf{0},\boldsymbol{\Sigma}')$ is implicitly assumed in Eq. (3.38).

3.6.1 Multiple experiments

Sometimes it is meaningful to consider a number of similar experiments. Even though the vector notation employed thus far implicitly permits their analysis already to some degree, e.g. see the matrix form of linear regression in the next section, it is revelatory to treat them in a more explicit fashion. Let us assume that a number of $n \in \mathbb{N}_{>0}$ different experiments are conducted, in each of which the collected data are explained with a vector-valued forward model as $\boldsymbol{y}_j = \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j) + \boldsymbol{\varepsilon}_j$. The unknowns $\boldsymbol{x}$ are constant throughout the experiments, whereas the experimental conditions $\boldsymbol{d}_j$ and covariances $\boldsymbol{\Sigma}_j$ may very well differ for $j = 1, \dots, n$. Assuming that the data in each experiment are conditionally independent given the unknowns, one has the data model

$$\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n | \boldsymbol{x} \sim \prod_{j=1}^{n} \mathcal{N}(\boldsymbol{y}_j | \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j), \boldsymbol{\Sigma}_j). \tag{3.39}$$

In connection with various types of model prediction errors and uncertainties, some generalizations of the described setup are discussed in Section 3.6.3. Against this background, Eq. (3.39) is rewritten as

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_n \end{pmatrix} | \boldsymbol{x} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} | \begin{pmatrix} \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_1) \\ \vdots \\ \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_n) \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 & \dots & \boldsymbol{0}_{N\times N} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0}_{N\times N} & \dots & \boldsymbol{\Sigma}_n \end{pmatrix} \right). \tag{3.40}$$

For the corresponding likelihood function one has $\mathcal{L}(\boldsymbol{x}) = \prod_{j=1}^{n} \mathcal{N}(\boldsymbol{y}_j | \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j), \boldsymbol{\Sigma}_j)$. The Bayesian update to the posterior density $\pi(\boldsymbol{x}|\boldsymbol{y}_1, \dots, \boldsymbol{y}_n) \propto \mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})$ and all further analyses proceed as normal.

3.6.2 Linear regression

In order to get a feel for inverse problems as discussed previously, we consider Bayesian linear regression. This can be viewed as the prototype of a linear inverse problem and often has a simple solution, i.e. when the prior and the likelihood are conjugate to each other. Details can be found in many references for Bayesian inference [109, 110] and regression analysis [111, 112]. Following a discussion of the linear model with known error variance, the natural extension with an unknown variance is investigated.

3.6.2.1 Known variance

In linear regression, the measured variables $\tilde{\boldsymbol{y}} \in \mathbb{R}^N$ are linearly dependent on a vector of unknown coefficients $\boldsymbol{x} \in \mathbb{R}^M$. The predictor-dependent design matrix $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ establishes this relation through $\tilde{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{x}$. A general linear regression model of the form as in Eq. (3.31) is thus written as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}. \tag{3.41}$$

The actual observations $\boldsymbol{y}$ are represented as a noisy version of the model outputs. As in Eq. (3.32), the noise is often modeled with a Gaussian distribution. We consider spherical errors with a diagonal covariance matrix of the form $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_N$, where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. They are jointly distributed according to

$$\pi(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon} | \boldsymbol{0}, \sigma^2 \boldsymbol{I}_N). \tag{3.42}$$

The probabilistic data model as in Eq. (3.33) is thus given as $\boldsymbol{Y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}_N)$ and the likelihood function in Eq. (3.34) follows as $\mathcal{L}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}_N)$.

As for the prior, one imposes a Gaussian distribution with a mean vector $\boldsymbol{\mu}_0$ and a covariance matrix $\boldsymbol{\Sigma}_0$ on the regressors. This is done by setting

$$\pi(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{3.43}$$

Again, the unknowns and the errors are treated as independent from one another. Given the linear-normal model in Eqs. (3.41) to (3.43), one can easily show that the posterior $\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})$ of the regression coefficients is the normal distribution

$$\pi(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad \text{with} \quad \begin{cases} \boldsymbol{\Sigma}_1 = \left(\boldsymbol{\Sigma}_0^{-1} + \sigma^{-2}\boldsymbol{A}^\top\boldsymbol{A}\right)^{-1}, \\ \boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \sigma^{-2}\boldsymbol{A}^\top\boldsymbol{y}\right). \end{cases} \tag{3.44}$$

Some useful intuition about the problem can be obtained from Eq. (3.44). On the one hand, in the limit of vanishing prior knowledge, i.e. $\boldsymbol{\Sigma}_0^{-1}$ can be neglected, the classical results $\boldsymbol{\mu}_1 = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y}$ and $\boldsymbol{\Sigma}_1 = \sigma^2(\boldsymbol{A}^\top\boldsymbol{A})^{-1}$ are obtained. On the other hand, for $\sigma^2 \to \infty$ one simply has $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0$, i.e. the prior knowledge dominates.

3.6.2.2 Unknown variance

A straightforward extension to the linear regression model with known variance is based on the inference of the error variance as an additional unknown. One considers an error model $\pi(\boldsymbol{\varepsilon}|\sigma^2) = \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$ with a covariance matrix of the form $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_N$. The errors for different response variables are uncorrelated and share the same variance just as in Eq. (3.42). Unlike before, however, the variance σ^2 is now treated as uncertain and associated with a prior distribution $\pi(\sigma^2)$. The latter forms a marginal of the joint prior $\pi(\boldsymbol{x}, \sigma^2)$ which is not necessarily the product $\pi(\boldsymbol{x}, \sigma^2) = \pi(\boldsymbol{x})\pi(\sigma^2)$. A statistical model $\boldsymbol{Y}|\boldsymbol{x}, \sigma^2 \sim \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}_N)$ gives rise to the likelihood function $\mathcal{L}(\boldsymbol{x}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}_N)$. The posterior $\pi(\boldsymbol{x}, \sigma^2|\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{x}, \sigma^2)\pi(\boldsymbol{x}, \sigma^2)$ is then taken as the basis for the identification of the unknowns $(\boldsymbol{x}, \sigma^2)$. One may extract $\pi(\boldsymbol{x}|\boldsymbol{y})$ and $\pi(\sigma^2|\boldsymbol{y})$ from the posterior by an appropriate marginalization as in Eq. (3.13). Before this formulation is simplified by using a conjugate prior for the variance σ^2, it is remarked that one may just as well assign a prior to the unknown standard deviation σ. See Section 3.5 for a discussion on related parametrization issues.

Similar to the precursory model with known variance, Bayesian linear regression with unknown variance possesses an explicit posterior presentation under a certain prior distribution. This conjugate prior has the hierarchical structure $\pi(\boldsymbol{x}, \sigma^2) = \pi(\boldsymbol{x}|\sigma^2)\pi(\sigma^2)$. The marginal prior of the variance $\sigma^2 > 0$ is an inverse gamma distribution $\pi(\sigma^2) = \mathcal{IG}(\sigma^2|\alpha_0, \beta_0) = \Gamma^{-1}(\alpha_0)\beta_0^{\alpha_0}(\sigma^2)^{-\alpha_0-1}\exp(-\beta_0/\sigma^2)$. Here, Γ is the gamma function and $\alpha_0, \beta_0 > 0$ are the shape and scale parameters of the distribution, respectively. The conditional prior of the unknown coefficients $\boldsymbol{x}$ is a normal distribution $\pi(\boldsymbol{x}|\sigma^2) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_0, \sigma^2\boldsymbol{\Sigma}_0)$ with the mean $\boldsymbol{\mu}_0$ and the σ^2-dependent covariance matrix $\sigma^2\boldsymbol{\Sigma}_0$. All in all, the joint prior is the normal inverse gamma distribution

$$\pi(\boldsymbol{x}, \sigma^2) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_0, \sigma^2\boldsymbol{\Sigma}_0)\,\mathcal{IG}(\sigma^2|\alpha_0, \beta_0) = \mathcal{NIG}(\boldsymbol{x}, \sigma^2|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \alpha_0, \beta_0). \tag{3.45}$$

Given this prior, a straightforward calculation yields the posterior density $\pi(\boldsymbol{x}, \sigma^2|\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{x}, \sigma^2)\pi(\boldsymbol{x}, \sigma^2)$ in a closed form. In particular, the posterior can be analytically expressed as

$$\pi(\boldsymbol{x}, \sigma^2|\boldsymbol{y}) = \mathcal{NIG}(\boldsymbol{x}, \sigma^2|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \alpha_1, \beta_1), \quad \text{with} \quad \begin{cases} \boldsymbol{\Sigma}_1 = \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{A}^\top\boldsymbol{A}\right)^{-1}, \\ \boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{A}^\top\boldsymbol{y}\right), \\ \alpha_1 = \alpha_0 + \dfrac{N}{2}, \\ \beta_1 = \beta_0 + \dfrac{1}{2}\left(\boldsymbol{y}^\top\boldsymbol{y} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1\boldsymbol{\mu}_1\right). \end{cases} \tag{3.46}$$

The posterior $\pi(\boldsymbol{x}, \sigma^2|\boldsymbol{y}) = \pi(\boldsymbol{x}|\sigma^2, \boldsymbol{y})\pi(\sigma^2|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \sigma^2\boldsymbol{\Sigma}_1)\,\mathcal{IG}(\sigma^2|\alpha_1, \beta_1)$ in Eq. (3.46) is of the same normal inverse gamma shape as the prior in Eq. (3.45).

3.6.3 Model uncertainties

In many experimental situations, at least some of the simplifying assumptions made cannot actually be justified. This concerns the idealizations formalized in Eqs. (3.31) to (3.33). While the inverse theory established so far focuses on epistemic parameter uncertainties, it woefully neglects various other forms of uncertainty. The treatment of parametric variability is indeed the main topic of Chapters 4 and 5. Bayesian probability lends itself to an analysis of other sources of uncertainty and error, too. A non-exhaustive synopsis of related methods is provided below. They are based on a refined representation and parametrization of the relation between forward model predictions and real measurement data.

There is indeed a plethora of methods for representing and handling error or uncertainty in the context of predictive modeling. This includes *multiplicative errors* in the model outputs $\boldsymbol{y} = \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}) \cdot (1 + \boldsymbol{\varepsilon})$ and *errors-in-variables*, i.e. the experimental conditions are only inexactly measured [113–115]. We do not further delve into these issues, but rather focus on additive measurement and modeling errors.

3.6.3.1 Random error calibration

It seems to be overly optimistic to start from the premise that the covariance matrix $\boldsymbol{\Sigma}$ in Eq. (3.32) is prespecified before the data analysis. In order to relax this limiting assumption, one can parametrize the error model conveniently and infer its parameters as additional unknowns during data analysis [116–118]. A simple example of this type of *error calibration* was already encountered in Section 3.6.2. Let the error model $\pi(\boldsymbol{\varepsilon}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ be parametrized by a possibly multivariate parameter vector $\boldsymbol{\theta}$. Note that this could encompass correlation parameters. Eliciting a prior $\pi(\boldsymbol{x}, \boldsymbol{\theta})$ and constructing a likelihood $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y} - \mathcal{M}(\boldsymbol{x})|\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ allows for conditioning on the data via $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}, \boldsymbol{\theta})$. The way the likelihood is written signalizes that the formulation can be generalized beyond Gaussian errors. A caveat is that the approach may require a fair bit of data, though.

3.6.3.2 Predictive model correction

While the error calibration as discussed above allows one to account for random sources of uncertainty, e.g. measurement noise, it disregards systematic inadequacies of the forward model [119, 120]. The question of *model discrepancy* is certainly tricky, though. It concerns the model building process as a whole. Let us consider an experimental scenario where the available predictive model shows considerable deficits but cannot be substituted with a better model. Moreover, the data are too few to learn alternative predictive models, but they are still numerous enough to awaken the hope for an in-depth error analysis. In this borderline situation, that lies exactly in between more model-centered uncertainty quantification and purely data-based machine learning, one can try to detect and quantify the modeling errors by comparing the predicted system outputs to real data. In turn, this can serve as a guide to model correction.

In the following, the multi-experiment setup from Section 3.6.1 is generalized in order to account for model discrepancy. As a first step, the discrepancy is either treated as an unknown constant or a random variable. In the next step, it is formulated as an unknown or even random function of the experimental conditions. Some of the mathematical details are omitted for the sake of readability. This concerns a few standard independence assumptions and the subtleties of certain basis representations.

To begin with, the simplest extension is grounded on thinking of the model discrepancy as an unknown constant $\boldsymbol{\delta} \in \mathbb{R}^N$. Accordingly, for $j = 1, \dots, n$ the actual measurements are represented as the sum $\boldsymbol{y}_j = \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j) + \boldsymbol{\delta} + \boldsymbol{\varepsilon}_j$ of the forward model response, the unknown offset and the random errors. It is remarked that the unknown $\boldsymbol{\delta}$ is fixed throughout all experiments. The data are then probabilistically described by

$$\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n | \boldsymbol{x}, \boldsymbol{\delta} \sim \prod_{j=1}^{n} \mathcal{N}(\boldsymbol{y}_j | \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j) + \boldsymbol{\delta}, \boldsymbol{\Sigma}_j). \tag{3.47}$$

The corresponding likelihood function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\delta})$ follows easily. One proceeds by imposing a prior $\pi(\boldsymbol{x}, \boldsymbol{\delta}) = \pi(\boldsymbol{x})\pi(\boldsymbol{\delta})$ and learns the unknowns $(\boldsymbol{x}, \boldsymbol{\delta})$ via the Bayesian update $\pi(\boldsymbol{x}, \boldsymbol{\delta} | \boldsymbol{y}_1, \dots, \boldsymbol{y}_n) \propto \mathcal{L}(\boldsymbol{x}, \boldsymbol{\delta})\pi(\boldsymbol{x}, \boldsymbol{\delta})$. Statistical identifiability or the lack thereof could be relevant issues at that.

Another representation of model error in various experiments is a number of random variables $\boldsymbol{\Delta}_j \sim \mathcal{N}(\boldsymbol{\delta}_j | \boldsymbol{\mu}_{\boldsymbol{\Delta}}, \boldsymbol{\Sigma}_{\boldsymbol{\Delta}})$. For a clear exposition, these variables are assumed to be Gaussian with unknown mean $\boldsymbol{\mu}_{\boldsymbol{\Delta}}$ and known covariance $\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}$. They take on different values $\boldsymbol{\Delta}_j = \boldsymbol{\delta}_j$ in each experiment. Accordingly, the data are represented as $\boldsymbol{y}_j = \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j) + \boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j$. The discrepancy term now randomly varies across the experiments. Conditionally on the unknowns $(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{\Delta}})$, the data are the sum of independent Gaussian variables and follow

$$\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n | \boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{\Delta}} \sim \prod_{j=1}^{n} \mathcal{N}(\boldsymbol{y}_j | \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j) + \boldsymbol{\mu}_{\boldsymbol{\Delta}}, \boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_{\boldsymbol{\Delta}}). \tag{3.48}$$

Inference of the unknowns proceeds as for the related model in Eq. (3.47). The likelihood function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{\Delta}})$ results from Eq. (3.48) and the joint prior is specified as $\pi(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{\Delta}}) = \pi(\boldsymbol{x})\pi(\boldsymbol{\mu}_{\boldsymbol{\Delta}})$. Hence, the posterior is $\pi(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{\Delta}} | \boldsymbol{y}_1, \dots, \boldsymbol{y}_n) \propto \mathcal{L}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{\Delta}})\pi(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{\Delta}})$. Note that the identification of the mean value $\boldsymbol{\mu}_{\boldsymbol{\Delta}}$ typically requires more than just a single experiment.

An advanced approach is to treat the systematic model discrepancy as an unknown function $\delta : \mathbb{R}^D \to \mathbb{R}^N$. It represents the model bias $\delta(\boldsymbol{d})$ as a function of the experimental conditions $\boldsymbol{d} \in \mathbb{R}^D$. For this reason, the error in an experiment $j \in \{1, \dots, n\}$ is a fixed yet unknown value $\delta(\boldsymbol{d}_j)$ rather than a random outcome. The data are consequentially modeled as $\boldsymbol{y}_j = \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j) + \delta(\boldsymbol{d}_j) + \boldsymbol{\varepsilon}_j$. Here, the discrepancy term effectively absorbs the systematic error components of the model predictions such that zero-mean errors are justified. The goal of the analysis is now to identify both the unknown model parameters as well as the discrepancy function. A prior model of the unknown function that is sloppily denoted as $\pi(\delta(\cdot))$ has to be established, e.g. by using an appropriate basis representation [121, 122] with prior distributions for the unknown coefficients. The likelihood function $\mathcal{L}(\boldsymbol{x}, \delta(\cdot))$ would rest on the statistical model

$$\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n | \boldsymbol{x}, \delta(\cdot) \sim \prod_{j=1}^{n} \mathcal{N}(\boldsymbol{y}_j | \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_j) + \delta(\boldsymbol{d}_j), \boldsymbol{\Sigma}_j). \tag{3.49}$$

A Bayesian analysis then informs about the unknowns $(\boldsymbol{x}, \delta(\cdot))$. Letting $\pi(\boldsymbol{x}, \delta(\cdot)) = \pi(\boldsymbol{x})\pi(\delta(\cdot))$ the posterior is $\pi(\boldsymbol{x}, \delta(\cdot) | \boldsymbol{y}_1, \dots, \boldsymbol{y}_n) \propto \mathcal{L}(\boldsymbol{x}, \delta(\cdot))\pi(\boldsymbol{x}, \delta(\cdot))$. After the calibration, one can eventually use the corrected model predictions $\tilde{\boldsymbol{y}}' = \mathcal{M}(\hat{\boldsymbol{x}}, \boldsymbol{d}') + \hat{\delta}(\boldsymbol{d}')$ in the interpolation regime, i.e. within the range of tested experimental conditions. The extrapolation beyond this range requires a good deal of courage and caution, though. Notice that the forward model $\mathcal{M}(\boldsymbol{x}, \boldsymbol{d})$ is not corrected as a function of the unknowns $\boldsymbol{x}$, which are fixed across the experiments. The model correction $\delta(\boldsymbol{d})$ rather relates to $\mathcal{M}(\boldsymbol{x}, \boldsymbol{d})$ as a function of the covariates $\boldsymbol{d}$, while the

parameter $\boldsymbol{x}$ attains its most plausible value. Bear in mind that data must be collected for various different experimental conditions in order to meaningfully characterize the structural error.

One could treat model discrepancy also as a random function, e.g. as an unknown realization of a stochastic process with priorly unknown hyperparameters. In an original formulation [123–125] and its multivariate generalizations [126, 127] both the simulator $\mathcal{M}(\boldsymbol{x}, \boldsymbol{d})$ and the discrepancy function $\delta(\boldsymbol{d})$ are nonparametrically represented as realizations of Gaussian processes. At the same time it is conditioned on the observational data and the experimental design. The hyperparameters of the mean $m\colon \mathbb{R}^D \to \mathbb{R}^N$ and the covariance function $c\colon \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{N\times N}$ are unknown themselves and therefore endowed with priors. In a rather vague manner this is hinted at by writing $\pi(m(\cdot))$ and $\pi(c(\cdot,\cdot))$. Given the unknowns $(\boldsymbol{x}, m(\cdot), c(\cdot,\cdot))$, finite collections of random variables are jointly normal under the aforementioned modeling assumptions. So are the data

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_n \end{pmatrix} | \boldsymbol{x}, m(\cdot), c(\cdot,\cdot) \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} | \begin{pmatrix} \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_1) \\ \vdots \\ \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_n) \end{pmatrix} + \begin{pmatrix} m(\boldsymbol{d}_1) \\ \vdots \\ m(\boldsymbol{d}_n) \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 + c(\boldsymbol{d}_1, \boldsymbol{d}_1) & \dots & c(\boldsymbol{d}_1, \boldsymbol{d}_n) \\ \vdots & \ddots & \vdots \\ c(\boldsymbol{d}_n, \boldsymbol{d}_1) & \dots & \boldsymbol{\Sigma}_n + c(\boldsymbol{d}_n, \boldsymbol{d}_n) \end{pmatrix} \right). \tag{3.50}$$

This can be compared to Eqs. (3.40) and (3.49). With the likelihood function $\mathcal{L}(\boldsymbol{x}, m(\cdot), c(\cdot,\cdot))$ that arises from Eq. (3.50) and the prior distribution $\pi(\boldsymbol{x}, m(\cdot), c(\cdot,\cdot)) = \pi(\boldsymbol{x})\pi(m(\cdot))\pi(c(\cdot,\cdot))$ one can construct the posterior $\pi(\boldsymbol{x}, m(\cdot), c(\cdot,\cdot)|\boldsymbol{y}_1, \dots, \boldsymbol{y}_n) \propto \mathcal{L}(\boldsymbol{x}, m(\cdot), c(\cdot,\cdot))\pi(\boldsymbol{x}, m(\cdot), c(\cdot,\cdot))$. The obtained results can be subsequently used for correcting model predictions and quantifying their uncertainty. This is elegant in theory and important in practice [128]. It allows for Bayesian model validation [129, 130] and for a coherent management of the uncertainties emerging from measurement errors, model inadequacies and a limited number of simulator runs [131]. On the downside, the approach is quite complex and identifiability may very well become a problem [132, 133]. Sufficiently numerous data subject to different experimental conditions have to be available.

3.6.3.3 Model class comparison

After the consideration of the uncertainties and discrepancies of a single predictive model, one can also assess the relative plausibilities of multiple alternative models. The evidence framework of Section 3.4 is eminently suitable for this kind of job. A number $L \in \mathbb{N}_{>1}$ of different Bayesian models $\boldsymbol{\mathcal{H}} = \{\mathcal{H}_1, \dots, \mathcal{H}_L\}$ is considered. They are referred to as *model classes* in common parlance. Each $\mathcal{H} \in \boldsymbol{\mathcal{H}}$ is characterized by a forward model $\mathcal{M}_\mathcal{H}$ with $D_\mathcal{H} \in \mathbb{N}_{>0}$ control variables $\boldsymbol{d}_\mathcal{H} \in \mathbb{R}^{D_\mathcal{H}}$ and $M_\mathcal{H} \in \mathbb{N}_{>0}$ unknown parameters $\boldsymbol{x}_\mathcal{H} \in \mathbb{R}^{M_\mathcal{H}}$, a prior distribution $\pi(\boldsymbol{x}_\mathcal{H}|\mathcal{H})$ and an error model $\pi(\boldsymbol{\varepsilon}_\mathcal{H}|\boldsymbol{\theta}_\mathcal{H}, \mathcal{H})$. The latter depends on model-specific parameters $\boldsymbol{\theta}_\mathcal{H}$ and might be associated with a prior $\pi(\boldsymbol{\theta}_\mathcal{H}|\mathcal{H})$. The plausibilities of different model classes can then be assessed by reference to their model evidences $Z_\mathcal{H}$ in Eq. (3.18). More generally, when a model prior $\pi(\mathcal{H})$ is available, the model posterior distribution $\pi(\mathcal{H}|\boldsymbol{y}) \propto Z_\mathcal{H}\pi(\mathcal{H})$ in Eq. (3.19) provides the basis for data-informed model class comparison, selection and averaging.

In inverse modeling, the outlined way of evaluating model classes can be deployed for comparing and selecting forward models [134–136]. Averaging over the model classes quantifies the prediction uncertainty in this context [137, 138]. It is worth noting that random error models can undergo the same procedure, too. For instance, one might be interested in different error correlation structures [139].

3.7 Bayesian computations

Posterior densities of very simple problems can sometimes be derived analytically. In the linear-Gaussian examples in Section 3.6.2 the posterior densities had the explicit expressions in Eqs. (3.44) and (3.46). This is a rare exception rather than a rule. Most often, the posterior density in Eq. (3.6) does not have such a closed-form solution. One therefore contents oneself with computational approximations of a few hand-picked characteristics of the full posterior. These typically take the form of the integrals in Eqs. (3.7) and (3.8) [140, 141].

Despite the restriction to posterior summaries only, computational Bayesian inference is still a challenging problem. Roughly speaking, any convenient algorithm oughts to master the difficulty that one only has limited access to the full posterior. For one thing, the posterior density cannot be evaluated directly, one can only compute the likelihood function for certain parameter values individually. Even if the densities of the prior or some auxiliary distribution can be calculated and sampled, this is a major constraint that all attempts have to cope with. For another thing, the computational budget may limit the number of calls to the likelihood which,

in turn, necessitates to select the corresponding parameter values wisely. This is especially important in inverse problems where the forward model needs to be run once for each likelihood evaluation.

Before presenting the most widespread approaches to computational Bayesian inference, i.e. random sampling and mathematical programming, it is remarked that researchers have recently devised a whole battery of new and creative alternatives. Among others, this includes hybrid schemes combining sampling and optimization [142, 143], importance sampling–based methods based on implicit transformations and Jacobian weights [144–147], rejection sampling–related techniques inspired by subset simulation for rare event estimation in structural reliability [148–150] and conditional expectation–focused polynomial chaos expansion filters [151–154]. In addition, a linear least squares technique for computing the actual posterior density along with the model evidence and the conditional expectations is presented in Chapter 8. Yet another novel approach based on random variable transformations is investigated in Chapter 9.

3.7.1 Markov chain Monte Carlo

An important class of algorithms for Bayesian inference rests on *Markov chain Monte Carlo* (MCMC) sampling [155]. The basic idea here is to construct a Markov chain that is suitable for sampling from the posterior distribution and for estimating conditional expectations accordingly. This only needs pointwise evaluations of the unnormalized posterior density and thus dispenses from computing the marginal likelihood. MCMC algorithms are therefore sufficient for single-model parameter estimation, though, for multi-model inference one has to search for more suitable methods [156–158].

Most MCMC techniques are based on the original *Metropolis–Hastings* (MH) *algorithm* [159, 160]. Since more detailed introductions are also found in most chapters of Part II, only a short summary of the algorithm is given. An ergodic Markov chain $(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots)$ over the prior support whose invariant distribution equals the posterior is realized as follows. The chain is initialized at a certain point $\boldsymbol{x}^{(1)} \in \mathbb{R}^M$, e.g. the prior expected value or a random draw from the prior. Given the current state of the Markov chain $\boldsymbol{x}^{(t)}$, one samples a candidate state $\boldsymbol{x}^{(\star)}$ from an auxiliary proposal distribution $p(\boldsymbol{x}^{(\star)}|\boldsymbol{x}^{(t)})$. This may depend on the current state. The proposed state is then accepted as the new state $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(\star)}$ with the probability

$$\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(\star)}) = \min\left(1, \frac{\pi(\boldsymbol{x}^{(\star)}|\boldsymbol{y})\, p(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(\star)})}{\pi(\boldsymbol{x}^{(t)}|\boldsymbol{y})\, p(\boldsymbol{x}^{(\star)}|\boldsymbol{x}^{(t)})}\right). \tag{3.51}$$

Otherwise the proposal is rejected and the chain remains in its state $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)}$. It is worth pointing out that the MH acceptance in Eq. (3.51) calls for posterior ratios, hence, only unnormalized densities have to be known. The acceptance step is tantamount to a correction required for sampling the posterior by drawing only from the proposal. After a total number of iterations $T \in \mathbb{N}_{>0}$ deemed sufficiently high, the conditional expectation in Eq. (3.8) can be approximated as

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] \approx \frac{1}{T}\sum_{t=1}^{T} h(\boldsymbol{x}^{(t)}). \tag{3.52}$$

The MH updating scheme specified above defines an appropriate Markov chain transition kernel. One can write the general probability of acceptance of a newly proposed state given $\boldsymbol{x}^{(t)}$ as

$$\alpha(\boldsymbol{x}^{(t)}) = \int_{\mathbb{R}^M} \alpha(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(\star)})\, p(\boldsymbol{x}^{(\star)}|\boldsymbol{x}^{(t)})\, \mathrm{d}\boldsymbol{x}^{(\star)}. \tag{3.53}$$

The probability of rejection is consequently given as $1 - \alpha(\boldsymbol{x}^{(t)})$. Let $\delta_{\boldsymbol{x}^{(t)}}$ denote the Dirac point mass at the current state. With Eqs. (3.51) and (3.53) the MH transition density from $\boldsymbol{x}^{(t)}$ to $\boldsymbol{x}^{(t+1)}$ can be written as

$$\mathcal{K}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)}) = \alpha(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)})\, p(\boldsymbol{x}^{(t+1)}|\boldsymbol{x}^{(t)}) + (1 - \alpha(\boldsymbol{x}^{(t)}))\, \delta_{\boldsymbol{x}^{(t)}}(\boldsymbol{x}^{(t+1)}). \tag{3.54}$$

It is easy to show that transition kernels of the form as in Eq. (3.54) satisfy microscopic time reversibility $\pi(\boldsymbol{x}^{(t)}|\boldsymbol{y})\, \mathcal{K}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)}) = \pi(\boldsymbol{x}^{(t+1)}|\boldsymbol{y})\, \mathcal{K}(\boldsymbol{x}^{(t+1)}, \boldsymbol{x}^{(t)})$. Macroscopic reversibility then automatically follows

$$\pi(\boldsymbol{x}^{(t+1)}|\boldsymbol{y}) = \int_{\mathbb{R}^M} \pi(\boldsymbol{x}^{(t)}|\boldsymbol{y})\, \mathcal{K}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)})\, \mathrm{d}\boldsymbol{x}^{(t)}. \tag{3.55}$$

Hence, the target posterior is the invariant distribution of the Markov chain. Together with the ergodicity of the Markov chain, Eq. (3.55) lays the foundation for the approximation in Eq. (3.52).

Because the MH algorithm provides a basis for sampling arbitrary probability distributions, it has played a pivotal historical role for computational Bayesian inference. Notwithstanding the above, it also suffers from some

major problems. These involve the serial correlation of the obtained posterior sample as well as the associated decrease in efficiency when compared to an independent sampling. The tuning of the proposal distribution is a tedious task that often requires trial and error. It should be added that it is not possible to unequivocally diagnose the convergence of the simulation and to stop it accordingly. Heuristic checks can be performed at the most, e.g. restarting the algorithm with various initializations and comparing the obtained results. Regardless of the convergence rate that is dimension-independent in theory, the abovementioned issues cause severe problems in practice. Successfully checking, tuning and executing the algorithm is quite an art. That a high number of likelihood evaluations are needed during these procedures makes matters even worse.

In the context of inverse modeling, where a call to the likelihood function triggers a run of the forward simulator, a straightforward way to accelerate Bayesian inversion via MCMC is the deployment of cheap surrogates. This includes metamodels based on polynomial chaoses [161, 162], Gaussian processes [163, 164] or neural networks [165, 166]. This possibility was already mentioned in Section 2.3. More generally, one can use a vast array of advanced MCMC samplers, e.g. adaptive variants [167, 168] or particle-based annealing/tempering schemes [169, 170]. A highly efficient gradient-driven MCMC sampler that performs updates in an augmented variable space is implemented in Chapter 5.

3.7.2 Variational inference

In this section we briefly discuss the basics of *variational Bayesian* (VB) *inference* [171]. Traditionally VB techniques were developed in probabilistic machine learning [172], but they can be also used for inverse problems [173–176]. VB inference establishes a deterministic alternative to stochastic MCMC sampling, where the posterior is computed in an optimization procedure. In particular, a member from a parametric family of distributions is chosen such that its resemblance to the target posterior is maximized.

In order to assess the closeness or distinctness of distributions, one has to decide on a formal means to compare them. The KL divergence that was already encountered in Eq. (3.17) is often chosen to that end. In the present context, it is thought of as a non-negative measure of the difference between two probability distributions that attains zero in the case of perfect coincidence. Since the KL divergence is non-symmetric, it is not a distance metric in the strict sense, though. The KL divergence $D_{\mathrm{KL}}(q\|\pi(\cdot|\boldsymbol{y}))$ of the target $\pi(\cdot|\boldsymbol{y})$ from a candidate density q is

$$D_{\mathrm{KL}}(q\|\pi(\cdot|\boldsymbol{y})) = \int_{\mathbb{R}^M} \log\left(\frac{q(\boldsymbol{x})}{\pi(\boldsymbol{x}|\boldsymbol{y})}\right) q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \log Z - \mathcal{F}(q). \tag{3.56}$$

As one can easily see, the divergence in Eq. (3.56) is the difference between the constant log-evidence $\log Z$ and the so-called *free energy*

$$\mathcal{F}(q) = \int_{\mathbb{R}^M} \log\left(\frac{\mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})}{q(\boldsymbol{x})}\right) q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{3.57}$$

If $q = \pi(\cdot|\boldsymbol{y})$ would exactly equal the posterior, one would have $D_{\mathrm{KL}}(q\|\pi(\cdot|\boldsymbol{y})) = 0$ and $\mathcal{F}(q) = \log Z$. More generally one has $D_{\mathrm{KL}}(q\|\pi(\cdot|\boldsymbol{y})) \geq 0$ and the free energy establishes a lower bound of the model evidence through $\log Z \geq \mathcal{F}(q)$. It is instructive to further decompose the free energy in the two ways

$$\mathcal{F}(q) = \int_{\mathbb{R}^M} \log(\mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x}))\, q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + H_{\mathrm{S}}(q) = \int_{\mathbb{R}^M} \log(\mathcal{L}(\boldsymbol{x}))\, q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - D_{\mathrm{KL}}(q\|\pi). \tag{3.58}$$

The first part of Eq. (3.58) happens to contain the Shannon entropy $H_{\mathrm{S}}(q) = -\int_{\mathbb{R}^M} \log(q(\boldsymbol{x}))\, q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, which we came across in Eq. (3.16) once before. The second equation is the variational variant of Eq. (3.21) according to which the Bayesian update can be interpreted as a compromise between the goodness of the fit to the data and the closeness to the prior, as measured by $\int_{\mathbb{R}^M} \log(\mathcal{L}(\boldsymbol{x}))\, q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ and $D_{\mathrm{KL}}(q\|\pi) = \int_{\mathbb{R}^M} \log(q(\boldsymbol{x})/\pi(\boldsymbol{x}))\, q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, respectively.

Given a parametric family $\mathcal{Q}$ of probability densities, one can try to find the density $q \in \mathcal{Q}$ that best approximates the posterior $q \approx \pi(\cdot|\boldsymbol{y})$ in the KL sense. In the light of Eq. (3.56), minimizing the relative entropy $D_{\mathrm{KL}}(q\|\pi(\cdot|\boldsymbol{y}))$ is equivalent to maximizing the free energy $\mathcal{F}(q)$. This means

$$q = \operatorname*{arg\,min}_{q^\star \in \mathcal{Q}} D_{\mathrm{KL}}(q^\star\|\pi(\cdot|\boldsymbol{y})) \quad \Leftrightarrow \quad q = \operatorname*{arg\,max}_{q^\star \in \mathcal{Q}} \mathcal{F}(q^\star). \tag{3.59}$$

That criterion minimizes the entropy gain or information loss that is incurred by using a parametric approximation of the posterior. All in all, a stochastic optimization problem has to be solved, in the sense that the extremum of an expectation under the candidate distribution is sought.

Notice that, as opposed to $D_{\mathrm{KL}}(q\|\pi(\cdot|\boldsymbol{y}))$ in Eq. (3.56), the free energy $\mathcal{F}(q)$ in Eq. (3.57) does not depend on the model evidence Z. Thus it can be evaluated and maximized without the need of computing the evidence. Also note that the asymmetry of the KL divergence motivates the choice of minimizing $D_{\mathrm{KL}}(q\|\pi(\cdot|\boldsymbol{y}))$ rather than $D_{\mathrm{KL}}(\pi(\cdot|\boldsymbol{y})\|q) = \int_{\mathbb{R}^M} \log(\pi(\boldsymbol{x}|\boldsymbol{y})/q(\boldsymbol{x}))\,\pi(\boldsymbol{x}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}$ which involves intractable posterior expectations. The reverse choice indeed leads to yet another algorithm [177].

Classically, in VB inference one considers factorized candidate distributions based on the mean field approximation [178]. This formulation can be enriched through copulas for representing multivariate dependencies [179, 180]. One may also contemplate the use of Gaussian mixture distributions [181, 182]. No matter what family of distributions is considered, the conditional expectation in Eq. (3.8) is, after the minimization in Eq. (3.59), evaluated as the correspondent average over the best posterior approximation

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] \approx \int_{\mathbb{R}^M} h(\boldsymbol{x})\,q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{3.60}$$

One may obtain corresponding results analytically for some simple QoIs. More frequently, the approximation in Eq. (3.60) has to be computed in a sampling-based procedure.

3.7.3 Laplace approximations

By running the algorithm longer and longer and drawing more and more samples, MCMC allows one to compute all relevant posterior summaries exactly in principle, i.e. it is asymptotically exact. In contrast, VB inference is an approximate method in that the achievable accuracy is limited by the used class of candidate distributions. If the actual posterior is similar to a Gaussian density, e.g. roughly unimodal, smooth and symmetric, it makes sense to approximate it accordingly. This could happen within the variational inference framework or, alternatively, by using Laplace approximations. The approach only requires the computation of the global maximum of the log–posterior density and the local second-order partial derivatives.

The normal approximation is sometimes motivated by results in asymptotic theory. Especially for well-specified data models it is sensible to study the asymptotic behavior of the posterior distribution in the large data sample limit from a frequentist point of view. Under suitable regularity conditions one can show asymptotic consistency of point estimators and asymptotic normality of the posterior distribution, see [183–185] for instance. Such results are sometimes referred to as the "Bayesian law of large numbers" and the "Bayesian central limit theorem". They provide theoretical insights and suggest Laplace approximations of the posterior. These approximations are based on the asymptotic analysis of integrals [186–189] and can be applied to the probability integrals of Bayesian inference [190–192] as well as uncertainty and reliability analysis [193–195].

Let us define $T(\boldsymbol{x}) = \log(\mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})) = \log\mathcal{L}(\boldsymbol{x}) + \log\pi(\boldsymbol{x})$ as the logarithm of the unnormalized posterior density. We assume that this function is at least twice continuously differentiable at a point $\boldsymbol{x}_0 \in \mathbb{R}^M$ and then consider its second-order Taylor approximation about that point. This approximation is given as $T(\boldsymbol{x}) \approx T(\boldsymbol{x}_0) + \boldsymbol{J}(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\top \boldsymbol{H}(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)$. Here, $\boldsymbol{J}(\boldsymbol{x}_0) = \partial T(\boldsymbol{x})/\partial\boldsymbol{x}^\top|_{\boldsymbol{x}_0} = (\partial T/\partial x_1(\boldsymbol{x}_0), \ldots, \partial T/\partial x_M(\boldsymbol{x}_0))$ is the gradient row-vector and $\boldsymbol{H}(\boldsymbol{x}_0) = \partial^2 T(\boldsymbol{x})/\partial\boldsymbol{x}\partial\boldsymbol{x}^\top|_{\boldsymbol{x}_0}$ with entries $H_{i,j}(\boldsymbol{x}_0) = \partial^2 T/\partial x_i \partial x_j(\boldsymbol{x}_0)$ for $i,j = 1,\ldots,M$ is the Hessian matrix of second-order partial derivatives. The corresponding Taylor series approximation around the posterior mode $\boldsymbol{x}_0 = \hat{\boldsymbol{x}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{x}\in\mathbb{R}^M} T(\boldsymbol{x})$, which is assumed to be the unique global maximum with a vanishing gradient $\boldsymbol{J}(\hat{\boldsymbol{x}}_{\mathrm{MAP}}) = \boldsymbol{0}$ and a generalized observed Fisher information matrix $-\boldsymbol{H}(\hat{\boldsymbol{x}}_{\mathrm{MAP}})$ that is positive definite, is then given as

$$T(\boldsymbol{x}) \approx T(\hat{\boldsymbol{x}}_{\mathrm{MAP}}) + \frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\mathrm{MAP}})^\top \boldsymbol{H}(\hat{\boldsymbol{x}}_{\mathrm{MAP}})(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\mathrm{MAP}}). \tag{3.61}$$

Based on Eq. (3.61) one can calculate approximations of the posterior density $\pi(\boldsymbol{x}|\boldsymbol{y})$ in Eq. (3.6), the model evidence Z in Eq. (3.7) and certain conditional expectation values $\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}]$ of the form as in Eq. (3.8). First of all, the so-called *Laplace approximation* of the posterior density $\pi(\boldsymbol{x}|\boldsymbol{y}) = Z^{-1}\exp(T(\boldsymbol{x}))$ is

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \approx \frac{\exp(T(\hat{\boldsymbol{x}}_{\mathrm{MAP}}))}{\hat{Z}} \exp\left(\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\mathrm{MAP}})^\top \boldsymbol{H}(\hat{\boldsymbol{x}}_{\mathrm{MAP}})(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\mathrm{MAP}})\right) = \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{x}}_{\mathrm{MAP}}, -\boldsymbol{H}^{-1}(\hat{\boldsymbol{x}}_{\mathrm{MAP}})). \tag{3.62}$$

That is a multivariate Gaussian density $\pi(\boldsymbol{x}|\boldsymbol{y}) \approx \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))/\sqrt{(2\pi)^M\det(\boldsymbol{\Sigma})}$ with mean $\boldsymbol{\mu} = \hat{\boldsymbol{x}}_{\mathrm{MAP}}$ and covariance matrix $\boldsymbol{\Sigma} = -\boldsymbol{H}^{-1}(\hat{\boldsymbol{x}}_{\mathrm{MAP}})$. Based on *Laplace's method* for integrals, the approximation of the model evidence $Z = \int_{\mathbb{R}^M}\exp(T(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}$ is given as the Gaussian integral

$$\begin{aligned} Z &\approx \exp(T(\hat{\boldsymbol{x}}_{\mathrm{MAP}})) \int_{\mathbb{R}^M} \exp\left(\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\mathrm{MAP}})^\top \boldsymbol{H}(\hat{\boldsymbol{x}}_{\mathrm{MAP}})(\boldsymbol{x}-\hat{\boldsymbol{x}}_{\mathrm{MAP}})\right)\mathrm{d}\boldsymbol{x} \\ &= \exp(T(\hat{\boldsymbol{x}}_{\mathrm{MAP}}))\sqrt{(2\pi)^M\det(-\boldsymbol{H}^{-1}(\hat{\boldsymbol{x}}_{\mathrm{MAP}}))} = \hat{Z}. \end{aligned} \tag{3.63}$$

This is the factor $\hat{Z}$ needed for the normalization of the distribution kernel in Eq. (3.62). The approximation in Eq. (3.63) is accurate if the posterior is strongly concentrated around its mode, such that the integral value is virtually dominated by the integrand at that point and its immediate vicinity.

The posterior expectation $\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \int_{\mathbb{R}^M} h(\boldsymbol{x})\mathcal{L}(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} / \int_{\mathbb{R}^M} \mathcal{L}(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ of a strictly positive and twice differentiable QoI $h\colon \mathbb{R}^M \to \mathbb{R}^+$ can be approximated by using Laplace's method separately for the numerator and denominator. To that end, we define $\check{T}(\boldsymbol{x}) = \log(h(\boldsymbol{x})\mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})) = \log h(\boldsymbol{x}) + \log \mathcal{L}(\boldsymbol{x}) + \log \pi(\boldsymbol{x})$ and assume that this function has a unique maximum at $\hat{\boldsymbol{x}}_{\max} = \arg\max_{\boldsymbol{x}\in\mathbb{R}^M} \check{T}(\boldsymbol{x})$. With a second-order Taylor expansion $\check{T}(\boldsymbol{x}) \approx \check{T}(\hat{\boldsymbol{x}}_{\max}) + \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_{\max})^\top \check{\boldsymbol{H}}(\hat{\boldsymbol{x}}_{\max})(\boldsymbol{x} - \hat{\boldsymbol{x}}_{\max})$ around the point $\hat{\boldsymbol{x}}_{\max}$, where $\check{\boldsymbol{H}}(\hat{\boldsymbol{x}}_{\max}) = \partial^2\check{T}(\boldsymbol{x})/\partial\boldsymbol{x}\partial\boldsymbol{x}^\top|_{\hat{\boldsymbol{x}}_{\max}}$ denotes the Hessian, we obtain the *fully exponential approximation*

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \frac{\int_{\mathbb{R}^M} \exp(\check{T}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^M} \exp(T(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}} \approx \frac{\exp(\check{T}(\hat{\boldsymbol{x}}_{\max}))}{\exp(T(\hat{\boldsymbol{x}}_{\mathrm{MAP}}))} \sqrt{\frac{\det(-\check{\boldsymbol{H}}^{-1}(\hat{\boldsymbol{x}}_{\max}))}{\det(-\boldsymbol{H}^{-1}(\hat{\boldsymbol{x}}_{\mathrm{MAP}}))}}. \tag{3.64}$$

Note that the strict positivity of $h(\boldsymbol{x})$ is required for the logarithm. Sometimes, one can add a large constant $c > 0$ to the function, such that $q(\boldsymbol{x}) = h(\boldsymbol{x}) + c$ is positive. The posterior expectation in Eq. (3.64) is then obtained from the approximation of $\mathbb{E}[q(\boldsymbol{X})|\boldsymbol{y}]$ by subtraction $\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \mathbb{E}[q(\boldsymbol{X})|\boldsymbol{y}] - c$.

Practically speaking, Laplace approximations merely call for locating the posterior mode and for computing the Hessian matrix of the unnormalized log-density at the mode. The mean of the normal approximation is then simply given by the posterior mode and the covariance matrix is the negative inverse Hessian. Numerical optimization and differentiation replace the original Bayesian integration problem. This is only slightly more difficult than simple MLE or MAP estimation and one can choose between a variety of standard optimizers, some of which approximate the Hessian anyhow, e.g. the Newton–Raphson method for optimization.

The procedure yields good approximations to posterior distributions that are roughly bell-shaped, i.e. unimodal and symmetric or at least strongly peaked and dominated by such a mode. Multimodal, highly skewed or heavy-tailed distributions cannot be captured well in general, though. Note that for a finite sample size the validity of the discussed normal approximation depends on the chosen parametrization. As opposed to variational inference where a global divergence measure is minimized, the Laplace approximation solely rests on local information such as the posterior mode and the curvature of the unnormalized log–posterior density. The Laplace approximation is thus generally easier to obtain. However, it can produce misleading results even in cases where variational inference with Gaussian candidates would still provide reasonable approximations, e.g. think of a double-peaked posterior with a dominating flat mode that accumulates most of the total probability mass and a peaked one that contains negligible mass but maximizes the density.

There are ways to widen the scope of applicability Laplace's method, that seems to be restricted to certain simple distributions only. These include Gaussian mixture approximations of multimodal distributions where modes, covariances and relative weights are computed for each mode [196], and iterative strategies for handling non-Gaussian mode shapes [197]. Beyond that, one can also use Laplace's method for obtaining a first crude posterior approximation in quick way. Once this approximation is at one's disposal, it can be refined at one's discretion. This could happen via importance sampling or a Metropolis–Hastings algorithm initialized at the posterior mode and driven by independent proposals sampled from the obtained approximation.

3.7.4 Log-likelihood function

A quite practical issue in Bayesian inference concerns the *log-likelihood function* $\log \mathcal{L}(\boldsymbol{x})$, i.e. the natural logarithm of the likelihood $\mathcal{L}(\boldsymbol{x})$. For both analytical and numerical approaches to statistical inference, it may be more convenient to work with the log-likelihood instead of the likelihood directly. This is due to the properties of the logarithm, i.e. strictly monotonic and smooth, and the typical structure of the likelihood, e.g. products with many factors. For instance, the likelihood and its logarithm may take the form

$$\mathcal{L}(\boldsymbol{x}) = \prod_{i=1}^{N} \pi(y_i|\boldsymbol{x}), \quad \log \mathcal{L}(\boldsymbol{x}) = \sum_{i=1}^{N} \log \pi(y_i|\boldsymbol{x}). \tag{3.65}$$

Note that individual likelihood terms $\pi(y_i|\boldsymbol{x})$ often contain exponentials anyway, e.g. as in the case of a Gaussian distribution. Beyond that, this happens whenever a member of the exponential family is used.

In frequentist inference one often tries to maximize the likelihood as in Eq. (3.3). Since the logarithm is monotonically increasing, this is equivalent to maximizing the logarithm, i.e. $\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}\in\mathbb{R}^M} \mathcal{L}(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}\in\mathbb{R}^M} \log \mathcal{L}(\boldsymbol{x})$. For finding the extrema one would have to zero the derivative of Eq. (3.65) with respect to the unknowns. Now, the explicit differentiation of a sum is easier than of a product, which is why the log-likelihood is preferable in an analytical analysis.

From a numerical perspective, taking the logarithm in intermediate steps of the computations may prevent from over- and underflow. If data are numerous and the likelihood in Eq. (3.65) contains many terms, the calculation in finite-precision floating-point arithmetic may easily suffer from that problem. A remedy is to use a common scaling factor. The problem can also be mitigated by performing the necessary manipulations as long as possible on the log-scale while hoping that problematic terms neutralize each other. In MCMC-based Bayesian inference one can defer the final exponentiation until the MH acceptance in Eq. (3.51) by

$$\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(\star)}) = \min\left\{1, \exp\left(\log \pi(\boldsymbol{x}^{(\star)}|\boldsymbol{y}) + \log p(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(\star)}) - \log \pi(\boldsymbol{x}^{(t)}|\boldsymbol{y}) - \log p(\boldsymbol{x}^{(\star)}|\boldsymbol{x}^{(t)})\right)\right\}. \tag{3.66}$$

When the likelihood is based on Gaussian distribution as in Eq. (3.34) with known covariance matrix $\boldsymbol{\Sigma}$, the normalization factor $((2\pi)^N \det(\boldsymbol{\Sigma}))^{-1/2}$ could be omitted from the MH acceptance probability. Otherwise, in case the covariance matrix and its determining parameters are arguments of the likelihood function, the log-determinant of the covariance matrix is usually calculated using a Cholesky factorization $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$ as $\log\det(\boldsymbol{\Sigma}) = \log\det(\boldsymbol{L}\boldsymbol{L}^\top) = 2\sum_{i=1}^{M} \log L_{ii}$. Here, $\boldsymbol{L}$ is a lower triangular matrix. This avoids the aforementioned numerical issues.

For the computation of the Bayes factor in Eq. (3.20) one typically deploys a reasonable scaling factor. Let $\mathcal{L}_{\mathcal{H}_1}(\boldsymbol{x}_{\mathcal{H}_1}) = \pi(\boldsymbol{y}|\boldsymbol{x}_{\mathcal{H}_1}, \mathcal{H}_1)$ and $\mathcal{L}_{\mathcal{H}_2}(\boldsymbol{x}_{\mathcal{H}_2}) = \pi(\boldsymbol{y}|\boldsymbol{x}_{\mathcal{H}_2}, \mathcal{H}_2)$ denote the likelihood functions of two competing models. One only needs a vague idea about the maximal values of the log-likelihoods $\Upsilon_{\mathcal{H}_1} \approx \max_{\boldsymbol{x}_{\mathcal{H}_1}}(\log \mathcal{L}_{\mathcal{H}_1}(\boldsymbol{x}_{\mathcal{H}_1}))$ and $\Upsilon_{\mathcal{H}_2} \approx \max_{\boldsymbol{x}_{\mathcal{H}_2}}(\log \mathcal{L}_{\mathcal{H}_2}(\boldsymbol{x}_{\mathcal{H}_2}))$ in order to calculate the Bayes factor as

$$B_{1,2} = \frac{Z_{\mathcal{H}_1}}{Z_{\mathcal{H}_2}} = \exp\left(\Upsilon_{\mathcal{H}_1} - \Upsilon_{\mathcal{H}_2}\right) \frac{\int_{\mathbb{R}^{M_{\mathcal{H}_1}}} \exp(\log(\mathcal{L}_{\mathcal{H}_1}(\boldsymbol{x}_{\mathcal{H}_1})) - \Upsilon_{\mathcal{H}_1})\, \pi(\boldsymbol{x}_{\mathcal{H}_1}|\mathcal{H}_1)\, \mathrm{d}\boldsymbol{x}_{\mathcal{H}_1}}{\int_{\mathbb{R}^{M_{\mathcal{H}_2}}} \exp(\log(\mathcal{L}_{\mathcal{H}_1}(\boldsymbol{x}_{\mathcal{H}_2})) - \Upsilon_{\mathcal{H}_2})\, \pi(\boldsymbol{x}_{\mathcal{H}_2}|\mathcal{H}_2)\, \mathrm{d}\boldsymbol{x}_{\mathcal{H}_2}}. \tag{3.67}$$

Of course, this normalization/renormalization procedure is not limited to Bayes factors. It can facilitate the numerical computation of other likelihood-based integrals, too.

References

[1] H. Jeffreys. *Theory of Probability*. 3rd ed. Oxford Classic Texts in the Physical Sciences. Oxford, UK: Oxford University Press, 1961.

[2] B. de Finetti. *Theory of Probability: A critical introductory treatment*. Ed. by A. Machi and A. Smith. 2 vols. Wiley Classics Library. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 1974–1975.

[3] E. T. Jaynes. *Probability Theory: The Logic of Science*. Ed. by G. L. Bretthorst. Cambridge, UK: Cambridge University Press, 2003. DOI: 10.1017/CBO9780511790423.

[4] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003.

[5] J.-P. Florens, M. Mouchart, and J.-M. Rolin. *Elements of Bayesian Statistics*. Monographs and Textbooks in Pure and Applied Mathematics. New York: Marcel Dekker, Inc., 1990.

[6] J. Shao. *Mathematical Statistics*. 2nd ed. Springer Texts in Statistics. New York: Springer, 2003. DOI: 10.1007/b97553.

[7] R. W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. New York: Springer, 2010. DOI: 10.1007/978-0-387-93839-4.

[8] G. R. Shorack. *Probability for Statisticians*. Springer Texts in Statistics. New York: Springer, 2000. DOI: 10.1007/b98901.

[9] K. B. Athreya and S. N. Lahiri. *Measure Theory and Probability Theory*. Springer Texts in Statistics. New York: Springer, 2006. DOI: 10.1007/978-0-387-35434-7.

[10] M. M. Rao. *Conditional Measures and Applications*. 2nd ed. Pure and Applied Mathematics. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2005. DOI: 10.1201/9781420027433.

[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. 3rd ed. Texts in Statistical Science. Boca Raton, Florida, USA: CRC Press, 2014.

[12] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Texts in Statistical Science. Boca Raton, Florida, USA: CRC Press, 2016.

[13] R. C. Aster, B. Borchers, and C. H. Thurber. *Parameter Estimation and Inverse Problems*. 2nd ed. Waltham, Massachusetts, USA: Academic Press, 2012. DOI: 10.1016/c2009-0-61134-x.

[14] N.-Z. Sun and A. Sun. *Model Calibration and Parameter Estimation: For Environmental and Water Resource Systems*. New York: Springer, 2015. DOI: 10.1007/978-1-4939-2323-6.

[15] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge, UK: Cambridge University Press, 2015. DOI: 10.1017/CBO9781107706804.

[16] K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation: A Mathematical Introduction*. Texts in Applied Mathematics 62. Cham, Switzerland: Springer International Publishing, 2015. DOI: 10.1007/978-3-319-20325-6.

[17] S. Jackman. *Bayesian Analysis for the Social Sciences*. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2009. DOI: 10.1002/9780470686621.

[18] J. Gill. *Bayesian Methods: A Social and Behavioral Sciences Approach*. 3rd ed. Statistics in the Social and Behavioral Sciences Series. Boca Raton, Florida, USA: CRC Press, 2015.

[19] W. von der Linden, V. Dose, and U. von Toussaint. *Bayesian Probability Theory: Applications in the Physical Sciences*. Cambridge, UK: Cambridge University Press, 2014. DOI: 10.1017/CBO9781139565608.

[20] S. Andreon and B. Weaver. *Bayesian Methods for the Physical Sciences: Learning from Examples in Astronomy and Physics*. Springer Series in Astrostatistics 4. Cham, Switzerland: Springer International Publishing, 2015. DOI: 10.1007/978-3-319-15287-5.

[21] K.-V. Yuen. *Bayesian Methods for Structural Dynamics and Civil Engineering*. Singapore: John Wiley & Sons (Asia) Pte Ltd, 2010. DOI: 10.1002/9780470824566.

[22] J. M. Nichols and K. D. Murphy. *Modeling and Estimation of Structural Damage*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2016. DOI: 10.1002/9781118776995.

[23] S. Wu, P. Angelikopoulos, C. Papadimitriou, R. Moser, and P. Koumoutsakos. "A hierarchical Bayesian framework for force field selection in molecular dynamics simulations". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 374.2060, 20150032 (2015), pp. 1–23. DOI: 10.1098/rsta.2015.0032.

[24] I. Behmanesh and B. Moaveni. "Accounting for environmental variability, modeling errors, and parameter estimation uncertainties in structural identification". In: *Journal of Sound and Vibration* 374 (2016), pp. 92–110. DOI: 10.1016/j.jsv.2016.03.022.

[25] J. Mullins and S. Mahadevan. "Bayesian Uncertainty Integration for Model Calibration, Validation, and Prediction". In: *Journal of Verification, Validation and Uncertainty Quantification* 1.1, 011006 (2016), pp. 1–10. DOI: 10.1115/1.4032371.

[26] S. Sankararaman and S. Mahadevan. "Integration of model verification, validation, and calibration for uncertainty quantification in engineering systems". In: *Reliability Engineering & System Safety* 138 (2015), pp. 194–209. DOI: 10.1016/j.ress.2015.01.023.

[27] C. Li and S. Mahadevan. "Role of calibration, validation, and relevance in multi-level uncertainty integration". In: *Reliability Engineering & System Safety* 148 (2016), pp. 32–43. DOI: 10.1016/j.ress.2015.11.013.

[28] P. Diaconis. "Bayesian numerical analysis". In: *Statistical decision theory and related topics IV: Volume 1*. Ed. by S. S. Gupta and J. O. Berger. New York: Springer-Verlag, 1988, pp. 163–175.

[29] A. O'Hagan. "Some Bayesian numerical analysis". In: *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*. Ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford, UK: Oxford University Press, 1992, pp. 345–363.

[30] P. Hennig, M. A. Osborne, and M. Girolami. "Probabilistic numerics and uncertainty in computations". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 471.2179, 20150142 (2015), pp. 1–17. DOI: 10.1098/rspa.2015.0142.

[31] A. O'Hagan. "Bayes–Hermite quadrature". In: *Journal of Statistical Planning and Inference* 29.3 (1991), pp. 245–260. DOI: 10.1016/0378-3758(91)90002-V.

[32] M. Kennedy. "Bayesian quadrature with non-normal approximating functions". In: *Statistics and Computing* 8.4 (1998), pp. 365–375. DOI: 10.1023/A:1008832824006.

[33] M. A. Osborne, D. Duvenaud, R. Garnett, C. E. Rasmussen, S. J. Roberts, and Z. Ghahramani. "Active Learning of Model Evidence Using Bayesian Quadrature". In: *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 46–54.

[34] F.-X. Briol, C. Oates, M. Girolami, and M. A. Osborne. "Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees". In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 1162–1170.

[35] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford, UK: Oxford University Press, 2001.

[36] C. A. Rohde. *Introductory Statistical Inference with the Likelihood Function*. Cham, Switzerland: Springer International Publishing, 2014. DOI: 10.1007/978-3-319-10461-4.

[37] B. M. Ayyub. *Elicitation of Expert Opinions for Uncertainty and Risks*. Boca Raton, Florida, USA: CRC Press, 2001. DOI: 10.1201/9781420040906.

[38] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. Statistics in Practice. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2006. DOI: 10.1002/0470033312.

[39] R. E. Kass and L. Wasserman. "The Selection of Prior Distributions by Formal Rules". In: *Journal of the American Statistical Association* 91.435 (1996), pp. 1343–1370. DOI: 10.1080/01621459.1996.10477003.

[40] M. Ghosh. "Objective Priors: An Introduction for Frequentists". In: *Statistical Science* 26.2 (2011), pp. 187–202. DOI: 10.1214/10-STS338.

[41] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4 (1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[42] E. T. Jaynes. "Prior Probabilities". In: *IEEE Transactions on Systems Science and Cybernetics* 4.3 (1968), pp. 227–241. DOI: 10.1109/TSSC.1968.300117.

[43] P. Diaconis and D. Ylvisaker. "Conjugate Priors for Exponential Families". In: *The Annals of Statistics* 7.2 (1979), pp. 269–281. DOI: 10.1214/aos/1176344611.

[44] D. R. Insua and F. Ruggeri, eds. *Robust Bayesian Analysis*. Lecture Notes in Statistics 152. New York: Springer, 2000. DOI: 10.1007/978-1-4612-1306-2.

[45] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[46] T. Park and G. Casella. "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686. DOI: 10.1198/016214508000000337.

[47] S. Ji, Y. Xue, and L. Carin. "Bayesian Compressive Sensing". In: *IEEE Transactions on Signal Processing* 56.6 (2008), pp. 2346–2356. DOI: 10.1109/TSP.2007.914345.

[48] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag, 1985. DOI: 10.1007/978-1-4757-4286-2.

[49] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. 2nd ed. Springer Texts in Statistics. New York: Springer, 2007. DOI: 10.1007/0-387-71599-1.

[50] D. V. Lindley. "On a Measure of the Information Provided by an Experiment". In: *The Annals of Mathematical Statistics* 27.4 (1956), pp. 986–1005. DOI: 10.1214/aoms/1177728069.

[51] D. V. Lindley. *Bayesian Statistics: A Review*. CBMS-NSF Regional Conference Series in Applied Mathematics 2. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 1972. DOI: 10.1137/1.9781611970654.

[52] J. M. Bernardo. "Expected Information as Expected Utility". In: *The Annals of Statistics* 7.3 (1979), pp. 686–690. DOI: 10.1214/aos/1176344689.

[53] K. Chaloner and I. Verdinelli. "Bayesian Experimental Design: A Review". In: *Statistical Science* 10.3 (1995), pp. 273–304. DOI: 10.1214/ss/1177009939.

[54] J. M. Bernardo. "Reference Posterior Distributions for Bayesian Inference". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (1979), pp. 113–147.

[55] J. O. Berger, J. M. Bernardo, and D. Sun. "The formal definition of reference priors". In: *The Annals of Statistics* 37.2 (2009), pp. 905–938. DOI: 10.1214/07-AOS587.

[56] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2nd ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2006. DOI: 10.1002/047174882X.

[57] R. M. Gray. *Entropy and Information Theory*. 2nd ed. New York: Springer, 2011. DOI: 10.1007/978-1-4419-7970-4.

[58] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[59] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.4 (1948), pp. 623–656. DOI: 10.1002/j.1538-7305.1948.tb00917.x.

[60] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. DOI: 10.1214/aoms/1177729694.

[61] S. Kullback. *Information Theory and Statistics*. Mineola, New York, USA: Dover Publications, Inc., 1968.

[62] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer, 2002. DOI: 10.1007/b97636.

[63] G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics 27. Cambridge, UK: Cambridge University Press, 2008. DOI: 10.1017/CBO9780511790485.

[64] T. Ando. *Bayesian Model Selection and Statistical Modeling*. Statistics: A Series of Textbooks and Monographs. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2010. DOI: 10.1201/EBK1439836149.

[65] D. Draper. "Assessment and Propagation of Model Uncertainty". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 45–97.

[66] M. Clyde and E. I. George. "Model Uncertainty". In: *Statistical Science* 19.1 (2004), pp. 81–94. DOI: 10.1214/088342304000000035.

[67] S. G. Walker. "Bayesian inference with misspecified models". In: *Journal of Statistical Planning and Inference* 143.10 (2013), pp. 1621–1633. DOI: 10.1016/j.jspi.2013.05.013.

[68] R. E. Kass. "Bayes Factors in Practice". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 42.5 (1993), pp. 551–560. DOI: 10.2307/2348679.

[69] R. E. Kass and A. E. Raftery. "Bayes Factors". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795. DOI: 10.1080/01621459.1995.10476572.

[70] D. J. C. MacKay. "Bayesian Interpolation". In: *Neural Computation* 4.3 (1992), pp. 415–447. DOI: 10.1162/neco.1992.4.3.415.

[71] D. J. C. MacKay. "Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks". In: *Network: Computation in Neural Systems* 6.3 (1995), pp. 469–505. DOI: 10.1088/0954-898X_6_3_011.

[72] A. Vehtari and J. Ojanen. "A survey of Bayesian predictive methods for model assessment, selection and comparison". In: *Statistics Surveys* 6 (2012), pp. 142–228. DOI: 10.1214/12-SS102.

[73] A. Gelman, J. Hwang, and A. Vehtari. "Understanding predictive information criteria for Bayesian models". In: *Statistics and Computing* 24.6 (2014), pp. 997–1016. DOI: 10.1007/s11222-013-9416-2.

[74] M. Muto and J. L. Beck. "Bayesian Updating and Model Class Selection for Hysteretic Structural Models Using Stochastic Simulation". In: *Journal of Vibration and Control* 14.1–2 (2008), pp. 7–34. DOI: 10.1177/1077546307079400.

[75] S. H. Cheung and J. L. Beck. "Calculation of Posterior Probabilities for Bayesian Model Class Assessment and Averaging from Posterior Samples Based on Dynamic System Data". In: *Computer-Aided Civil and Infrastructure Engineering* 25.5 (2010), pp. 304–321. DOI: 10.1111/j.1467-8667.2009.00642.x.

[76] A. E. Raftery, D. Madigan, and J. A. Hoeting. "Bayesian Model Averaging for Linear Regression Models". In: *Journal of the American Statistical Association* 92.437 (1997), pp. 179–191. DOI: 10.1080/01621459.1997.10473615.

[77] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. "Bayesian Model Averaging: A Tutorial". In: *Statistical Science* 14.4 (1999), pp. 382–417. DOI: 10.1214/ss/1009212519.

[78] G. E. P. Box and P. W. Tidwell. "Transformation of the Independent Variables". In: *Technometrics* 4.4 (1962), pp. 531–550. DOI: 10.1080/00401706.1962.10490038.

[79] G. E. P. Box and D. R. Cox. "An Analysis of Transformations". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2 (1964), pp. 211–252.

[80] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Reading, Massachusetts, USA: Addison-Wesley Publishing Company, Inc., 1973. DOI: 10.1002/9781118033197.

[81] H. L. Harney. *Bayesian Inference: Data Evaluation and Decisions*. 2nd ed. Cham, Switzerland: Springer International Publishing, 2016. DOI: 10.1007/978-3-319-41644-1.

[82] H. Jeffreys. "An Invariant Form for the Prior Probability in Estimation Problems". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 186.1007 (1946), pp. 453–461. DOI: 10.1098/rspa.1946.0056.

[83] V. Isakov. *Inverse Problems for Partial Differential Equations*. 2nd ed. Applied Mathematical Sciences 127. New York: Springer, 2006. DOI: 10.1007/0-387-32183-7.

[84] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. 2nd ed. Applied Mathematical Sciences 120. New York: Springer, 2011. DOI: 10.1007/978-1-4419-8474-6.

[85] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2005. DOI: 10.1137/1.9780898717921.

[86] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Applied Mathematical Sciences 160. New York: Springer, 2005. DOI: 10.1007/b138659.

[87] W. Menke. *Geophysical Data Analysis: Discrete Inverse Theory*. 3rd ed. Waltham, Massachusetts, USA: Academic Press, 2012.

[88] M. S. Zhdanov. *Inverse Theory and Applications in Geophysics*. 2nd ed. Amsterdam, Netherlands: Elsevier, 2015. DOI: 10.1016/c2012-0-03334-0.

[89] A. F. Bennett. *Inverse Methods in Physical Oceanography*. Cambridge Monographs on Mechanics. Cambridge, UK: Cambridge University Press, 1992. DOI: 10.1017/CBO9780511600807.

[90] A. F. Bennett. *Inverse Modeling of the Ocean and Atmosphere*. Cambridge, UK: Cambridge University Press, 2002. DOI: 10.1017/CBO9780511535895.

[91] P. Boccacci and M. Bertero. *Introduction to Inverse Problems in Imaging*. Bristol, UK: IOP Publishing Ltd, 1998. DOI: 10.1201/9781439822067.

[92] B. Chalmond. *Modeling and Inverse Problems in Imaging Analysis*. Applied Mathematical Sciences 155. New York: Springer-Verlag, 2003. DOI: 10.1007/978-0-387-21662-1.

[93] K. I. Hopcraft and P. R. Smith. *An Introduction to Electromagnetic Inverse Scattering*. Developments in Electromagnetic Theory and Applications 7. Dordrecht, Netherlands: Springer, 1992. DOI: 10.1007/978-94-015-8014-4.

[94] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. 3rd ed. Applied Mathematical Sciences 93. New York: Springer, 2013. DOI: 10.1007/978-1-4614-4942-3.

[95] O. M. Alifanov. *Inverse Heat Transfer Problems*. International Series in Heat and Mass Transfer. Springer-Verlag Berlin Heidelberg, 1994. DOI: 10.1007/978-3-642-76436-3.

[96] M. N. Özisik and H. R. B. Orlande. *Inverse Heat Transfer: Fundamentals and Applications*. New York: Taylor & Francis, 2000.

[97] G. E. Stavroulakis. *Inverse and Crack Identification Problems in Engineering Mechanics*. Applied Optimization 46. Dordrecht, Netherlands: Springer, 2001. DOI: 10.1007/978-1-4615-0019-3.

[98] G. R. Liu and X. Han. *Computational Inverse Techniques in Nondestructive Evaluation*. Boca Raton, Florida, USA: CRC Press, 2003. DOI: 10.1201/9780203494486.

[99] G. M. L. Gladwell. *Inverse Problems in Vibration*. 2nd ed. Solid Mechanics and Its Applications 119. Dordrecht, Netherlands: Kluwer Academic Publishers, 2004. DOI: 10.1007/1-4020-2721-4.

[100] M. I. Friswell and J. E. Mottershead. *Finite Element Model Updating in Structural Dynamics*. Solid Mechanics and its Applications 38. Dordrecht, Netherlands: Springer, 1995. DOI: 10.1007/978-94-015-8508-8.

[101] T. Marwala. *Finite-element-model Updating Using Computional Intelligence Techniques: Applications to Structural Dynamics*. London, UK: Springer-Verlag, 2010. DOI: 10.1007/978-1-84996-323-7.

[102] T. Marwala, I. Boulkaibet, and S. Adhikari. *Probabilistic Finite Element Model Updating Using Bayesian Statistics: Applications to Aeronautical and Mechanical Engineering*. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2017.

[103] A. M. Stuart. "Inverse problems: A Bayesian perspective". In: *Acta Numerica* 19 (2010), pp. 451–559. DOI: 10.1017/S0962492910000061.

[104] M. Dashti and A. M. Stuart. "The Bayesian Approach to Inverse Problems". In: *Handbook of Uncertainty Quantification*. Ed. by R. Ghanem, D. Higdon, and H. Owhadi. Cham, Switzerland: Springer International Publishing, 2016. DOI: 10.1007/978-3-319-11259-6_7-1.

[105] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-posed Problems*. Scripta Series in Mathematics. Washington, D.C., USA: Winston & Sons, 1977.

[106] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola. *Numerical Methods for the Solution of Ill-Posed Problems*. Mathematics and Its Applications 328. Springer, 1995. DOI: 10.1007/978-94-015-8480-7.

[107] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications 375. Dordrecht, Netherlands: Kluwer Academic Publishers, 1996.

[108] J. Idier, ed. *Bayesian Approach to Inverse Problems*. Digital Signal and Image Processing Series. London, UK: ISTE Ltd, 2008. DOI: 10.1002/9780470611197.

[109] A. O'Hagan. *Bayesian Inference*. Kendall's Advanced Theory of Statistics 2B. London, UK: Edward Arnold Publishers Ltd, 1994.

[110] P. D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. New York: Springer, 2009. DOI: 10.1007/978-0-387-92407-6.

[111] G. A. F. Seber and A. J. Lee. *Linear Regression Analysis*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2003. DOI: 10.1002/9780471722199.

[112] X. Yan and X. G. Su. *Linear Regression Analysis: Theory and Computing*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2009. DOI: 10.1142/9789812834119.

[113] W. A. Fuller. *Measurement Error Models*. Wiley Series in Probability and Mathematical Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 1987. DOI: 10.1002/9780470316665.

[114] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Monographs on Statistics an Applied Probability 105. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2006.

[115] J. P. Buonaccorsi. *Measurement Error: Models, Methods, and Applications*. Interdisciplinary Statistics Series. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2010. DOI: 10.1201/9781420066586.

[116] E. Zhang, P. Feissel, and J. Antoni. "Bayesian model updating with consideration of modeling error". In: *European Journal of Computational Mechanics* 19.1–3 (2010), pp. 255–266. DOI: 10.3166/ejcm.19.255-266.

[117] E. L. Zhang, P. Feissel, and J. Antoni. "A comprehensive Bayesian approach for model updating and quantification of modeling errors". In: *Probabilistic Engineering Mechanics* 26.4 (2011), pp. 550–560. DOI: 10.1016/j.probengmech.2011.07.001.

[118] E. Zhang, J. Antoni, and P. Feissel. "Bayesian force reconstruction with an uncertain model". In: *Journal of Sound and Vibration* 331.4 (2012), pp. 798–814. DOI: 10.1016/j.jsv.2011.10.021.

[119] Y. Ling, J. Mullins, and S. Mahadevan. "Selection of model discrepancy priors in Bayesian calibration". In: *Journal of Computational Physics* 276 (2014), pp. 665–680. DOI: 10.1016/j.jcp.2014.08.005.

[120] K. Sargsyan, H. N. Najm, and R. Ghanem. "On the Statistical Calibration of Physical Models". In: *International Journal of Chemical Kinetics* 47.4 (2015), pp. 246–276. DOI: 10.1002/kin.20906.

[121] I. Farajpour and S. Atamturktur. "Error and Uncertainty Analysis of Inexact and Imprecise Computer Models". In: *Journal of Computing in Civil Engineering* 27.4 (2013), pp. 407–418. DOI: 10.1061/(ASCE)CP.1943-5487.0000233.

[122] Y. He and D. Xiu. "Numerical strategy for model correction using physical constraints". In: *Journal of Computational Physics* 313 (2016), pp. 617–634. DOI: 10.1016/j.jcp.2016.02.054.

[123] M. C. Kennedy and A. O'Hagan. "Bayesian calibration of computer models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 425–464. DOI: 10.1111/1467-9868.00294.

[124] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. "Combining Field Data and Computer Simulations for Calibration and Prediction". In: *SIAM Journal on Scientific Computing* 26.2 (2004), pp. 448–466. DOI: 10.1137/S1064827503426693.

[125] D. Higdon, J. D. McDonnell, N. Schunck, J. Sarich, and S. M. Wild. "A Bayesian approach for parameter estimation and prediction using a computationally intensive model". In: *Journal of Physics G: Nuclear and Particle Physics* 42.3, 034009 (2015), pp. 1–18. DOI: 10.1088/0954-3899/42/3/034009.

[126] D. Higdon, J. Gattiker, B. Williams, and M. Rightley. "Computer Model Calibration Using High-Dimensional Output". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 570–583. DOI: 10.1198/016214507000000888.

[127] D. Higdon, C. Nakhleh, J. Gattiker, and B. Williams. "A Bayesian calibration approach to the thermal problem". In: *Computer Methods in Applied Mechanics and Engineering* 197.29–32 (2008), pp. 2431–2441. DOI: 10.1016/j.cma.2007.05.031.

[128] J. Brynjarsdóttir and A. O'Hagan. "Learning about physical parameters: the importance of model discrepancy". In: *Inverse Problems* 30.11, 114007 (2014), pp. 1–24. DOI: 10.1088/0266-5611/30/11/114007.

[129] M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu. "A Framework for Validation of Computer Models". In: *Technometrics* 49.2 (2007), pp. 138–154. DOI: 10.1198/004017007000000092.

[130] S. Wang, W. Chen, and K.-L. Tsui. "Bayesian Validation of Computer Models". In: *Technometrics* 51.4 (2009), pp. 439–451. DOI: 10.1198/TECH.2009.07011.

[131] I. Bilionis and N. Zabaras. "Solution of inverse problems with limited forward solver evaluations: a Bayesian perspective". In: *Inverse Problems* 30.1, 015004 (2014), pp. 1–32. DOI: 10.1088/0266-5611/30/1/015004.

[132] P. D. Arendt, D. W. Apley, and W. Chen. "Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability". In: *Journal of Mechanical Design* 134.10, 100908 (2012), pp. 1–12. DOI: 10.1115/1.4007390.

[133] P. D. Arendt, D. W. Apley, W. Chen, D. Lamb, and D. Gorsich. "Improving Identifiability in Model Calibration Using Multiple Responses". In: *Journal of Mechanical Design* 134.10, 100909 (2012), pp. 1–9. DOI: 10.1115/1.4007573.

[134] L. Mthembu, T. Marwala, M. I. Friswell, and S. Adhikari. "Model selection in finite element model updating using the Bayesian evidence statistic". In: *Mechanical Systems and Signal Processing* 25.7 (2011), pp. 2399–2412. DOI: 10.1016/j.ymssp.2011.04.001.

[135] M. E. Riley and R. V. Grandhi. "Quantification of model-form and predictive uncertainty for multi-physics simulation". In: *Computers & Structures* 89.11–12 (2011), pp. 1206–1213. DOI: 10.1016/j.compstruc.2010.10.004.

[136] M. E. Riley and R. V. Grandhi. "Quantification of Modeling-Induced Uncertainties in Simulation-Based Analyses". In: *AIAA Journal* 52.1 (2014), pp. 195–202. DOI: 10.2514/1.J052871.

[137] I. Park, H. K. Amarchinta, and R. V. Grandhi. "A Bayesian approach for quantification of model uncertainty". In: *Reliability Engineering & System Safety* 95.7 (2010), pp. 777–785. DOI: 10.1016/j.ress.2010.02.015.

[138] I. Park and R. V. Grandhi. "Quantifying Multiple Types of Uncertainty in Physics-Based Simulation Using Bayesian Model Averaging". In: *AIAA Journal* 49.5 (2011), pp. 1038–1045. DOI: 10.2514/1.J050741.

[139] E. Simoen, C. Papadimitriou, and G. Lombaert. "On prediction error correlation in Bayesian model updating". In: *Journal of Sound and Vibration* 332.18 (2013), pp. 4136–4152. DOI: 10.1016/j.jsv.2013.03.019.

[140] M. Evans and T. Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford Statistical Science Series 20. Oxford, UK: Oxford University Press, 2000.

[141] M.-H. Chen, Q.-M. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer Series in Statistics. New York: Springer-Verlag, 2000. DOI: 10.1007/978-1-4612-1276-8.

[142] J. M. Bardsley, A. Solonen, H. Haario, and M. Laine. "Randomize-Then-Optimize: A Method for Sampling from Posterior Distributions in Nonlinear Inverse Problems". In: *SIAM Journal on Scientific Computing*. A 36.4 (2014), pp. 1895–1910. DOI: 10.1137/140964023.

[143] J. M. Bardsley, A. Seppänen, A. Solonen, H. Haario, and J. Kaipio. "Randomize-Then-Optimize for Sampling and Uncertainty Quantification in Electrical Impedance Tomography". In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1 (2015), pp. 1136–1158. DOI: 10.1137/140978272.

[144] A. J. Chorin and X. Tu. "Implicit sampling for particle filters". In: *Proceedings of the National Academy of Sciences* 106.41 (2009), pp. 17249–17254. DOI: 10.1073/pnas.0909196106.

[145] A. Chorin, M. Morzfeld, and X. Tu. "Implicit particle filters for data assimilation". In: *Communications in Applied Mathematics and Computational Science* 5.2 (2010), pp. 221–240. DOI: 10.2140/camcos.2010.5.221.

[146] A. J. Chorin and X. Tu. "An iterative implementation of the implicit nonlinear filter". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 46.3 (2012), pp. 535–543. DOI: 10.1051/m2an/2011055.

[147] M. Morzfeld, X. Tu, J. Wilkening, and A. J. Chorin. "Parameter estimation by implicit sampling". In: *Communications in Applied Mathematics and Computational Science* 10.2 (2015), pp. 205–225. DOI: 10.2140/camcos.2015.10.205.

[148] W. Betz, C. M. Mok, I. Papaioannou, and D. Straub. "Bayesian model calibration using structural reliability methods: Application to the hydrological abc model". In: *2nd International Conference on Vulnerability and Risk Analysis and Management and 6th International Symposium on Uncertainty Modeling and Analysis (ICVRAM & ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 276, pp. 2734–2743. DOI: 10.1061/9780784413609.276.

[149] D. Straub and I. Papaioannou. "Bayesian Updating with Structural Reliability Methods". In: *Journal of Engineering Mechanics* 141.3, 04014134 (2015), pp. 1–13. DOI: 10.1061/(ASCE)EM.1943-7889.0000839.

[150] F. A. DiazDelaO, A. Garbuno-Inigo, S. K. Au, and I. Yoshida. "Bayesian updating and model class selection with Subset Simulation". In: *Computer Methods in Applied Mechanics and Engineering* 317 (2017), pp. 1102–1121. DOI: 10.1016/j.cma.2017.01.006.

[151] B. V. Rosić, A. Litvinenko, O. Pajonk, and H. G. Matthies. "Sampling-free linear Bayesian update of polynomial chaos representations". In: *Journal of Computational Physics* 231.17 (2012), pp. 5761–5787. DOI: 10.1016/j.jcp.2012.04.044.

[152] B. V. Rosić, A. Kučerová, J. Sýkora, O. Pajonk, A. Litvinenko, and H. G. Matthies. "Parameter identification in a probabilistic setting". In: *Engineering Structures* 50 (2013), pp. 179–196. DOI: 10.1016/j.engstruct.2012.12.029.

[153] H. G. Matthies, E. Zander, B. V. Rosić, A. Litvinenko, and O. Pajonk. "Inverse Problems in a Bayesian Setting". In: *Computational Methods for Solids and Fluids: Multiscale Analysis, Probability Aspects and Model Reduction*. Ed. by A. Ibrahimbegovic. Vol. 41. Computational Methods in Applied Sciences. Cham, Switzerland: Springer International Publishing, 2016, pp. 245–286. DOI: 10.1007/978-3-319-27996-1_10.

[154] H. G. Matthies, E. Zander, B. V. Rosić, and A. Litvinenko. "Parameter estimation via conditional expectation: a Bayesian inversion". In: *Advanced Modeling and Simulation in Engineering Sciences* 3, 24 (2016), pp. 1–21. DOI: 10.1186/s40323-016-0075-7.

[155] D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd ed. Texts in Statistical Science 68. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2006.

[156] N. Friel and J. Wyse. "Estimating the evidence – a review". In: *Statistica Neerlandica* 66.3 (2012), pp. 288–308. DOI: 10.1111/j.1467-9574.2011.00515.x.

[157] A. Schöniger, T. Wöhling, L. Samaniego, and W. Nowak. "Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence". In: *Water Resources Research* 50.12 (2014), pp. 9484–9513. DOI: 10.1002/2014WR016062.

[158] K. H. Knuth, M. Habeck, N. K. Malakar, A. M. Mubeen, and B. Placek. "Bayesian evidence and model selection". In: *Digital Signal Processing* 47 (2015), pp. 50–67. DOI: 10.1016/j.dsp.2015.06.012.

[159] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[160] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97.

[161] P. M. Tagade and H.-L. Choi. "A Generalized Polynomial Chaos-Based Method for Efficient Bayesian Calibration of Uncertain Computational Models". In: *Inverse Problems in Science and Engineering* 22.4 (2014), pp. 602–624. DOI: 10.1080/17415977.2013.823411.

[162] I. Sraj, O. P. Le Maître, O. M. Knio, and I. Hoteit. "Coordinate transformation and Polynomial Chaos for the Bayesian inference of a Gaussian process with parametrized prior covariance function". In: *Computer Methods in Applied Mechanics and Engineering* 298 (2016), pp. 205–228. DOI: 10.1016/j.cma.2015.10.002.

[163] P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos. "X-TMCMC: Adaptive kriging for Bayesian inverse modeling". In: *Computer Methods in Applied Mechanics and Engineering* 289 (2015), pp. 409–428. DOI: 10.1016/j.cma.2015.01.015.

[164] H.-P. Wan and W.-X. Ren. "Stochastic model updating utilizing Bayesian approach and Gaussian process model". In: *Mechanical Systems and Signal Processing* 70–71 (2016), pp. 245–268. DOI: 10.1016/j.ymssp.2015.08.011.

[165] C. Balaji and T. Padhi. "A new ANN driven MCMC method for multi-parameter estimation in two-dimensional conduction with heat generation". In: *International Journal of Heat and Mass Transfer* 53.23–24 (2010), pp. 5440–5455. DOI: 10.1016/j.ijheatmasstransfer.2010.05.064.

[166] T. Hauser, A. Keats, and L. Tarasov. "Artificial neural network assisted Bayesian calibration of climate models". In: *Climate Dynamics* 39.1 (2012), pp. 137–154. DOI: 10.1007/s00382-011-1168-0.

[167] H. Haario, E. Saksman, and J. Tamminen. "An adaptive Metropolis algorithm". In: *Bernoulli* 7.2 (2001), pp. 223–242. DOI: 10.2307/3318737.

[168] C. Andrieu and J. Thoms. "A tutorial on adaptive MCMC". In: *Statistics and Computing* 18.4 (2008), pp. 343–373. DOI: 10.1007/s11222-008-9110-y.

[169] J. Ching and Y. Chen. "Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection, and Model Averaging". In: *Journal of Engineering Mechanics* 133.7 (2007), pp. 816–832. DOI: 10.1061/(ASCE)0733-9399(2007)133:7(816).

[170] W. Betz, I. Papaioannou, and D. Straub. "Transitional Markov Chain Monte Carlo: Observations and Improvements". In: *Journal of Engineering Mechanics* 142.5, 04016016 (2016), pp. 1–10. DOI: 10.1061/(ASCE)EM.1943-7889.0001066.

[171] V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing*. Signals and Communication Technology. Springer-Verlag Berlin Heidelberg, 2006. DOI: 10.1007/3-540-28820-1.

[172] M. J. Wainwright and M. I. Jordan. "Graphical Models, Exponential Families, and Variational Inference". In: *Foundations and Trends in Machine Learning* 1.1–2 (2008), pp. 1–305. DOI: 10.1561/2200000001.

[173] B. Jin. "A variational Bayesian method to inverse problems with impulsive noise". In: *Journal of Computational Physics* 231.2 (2012), pp. 423–435. DOI: 10.1016/j.jcp.2011.09.009.

[174] N. Guha, X. Wu, Y. Efendiev, B. Jin, and B. K. Mallick. "A variational Bayesian approach for inverse problems with skew-t error distributions". In: *Journal of Computational Physics* 301 (2015), pp. 377–393. DOI: 10.1016/j.jcp.2015.07.062.

[175] I. M. Franck and P. S. Koutsourelakis. "Sparse Variational Bayesian approximations for nonlinear inverse problems: Applications in nonlinear elastography". In: *Computer Methods in Applied Mechanics and Engineering* 299 (2016), pp. 215–244. DOI: 10.1016/j.cma.2015.10.015.

[176] I. M. Franck and P. S. Koutsourelakis. "Multimodal, high-dimensional, model-based, Bayesian inverse problems with applications in biomechanics". In: *Journal of Computational Physics* 329 (2017), pp. 91–125. DOI: 10.1016/j.jcp.2016.10.039.

[177] T. P. Minka. "Expectation Propagation for Approximate Bayesian Inference". In: *17th Conference on Uncertainty in Artificial Intelligence (UAI 2001)*. Ed. by J. Breese and D. Koller. San Francisco, California, USA: Morgan Kaufmann Publishers, 2001, pp. 362–369.

[178] M. Opper and D. Saad, eds. *Advanced Mean Field Methods: Theory and Practice*. Neural Information Processing Series. Cambridge, Massachusetts, USA: The MIT Press, 2001.

[179] D. Tran, D. Blei, and E. M. Airoldi. "Copula variational inference". In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 3564–3572.

[180] S. Han, X. Liao, D. B. Dunson, and L. Carin. "Variational Gaussian Copula Inference". In: *JMLR Workshop and Conference Proceedings: 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016)* 51 (2016), pp. 829–838.

[181] O. Zobay. "Variational Bayesian inference with Gaussian-mixture approximations". In: *Electronic Journal of Statistics* 8.1 (2014), pp. 355–389. DOI: 10.1214/14-EJS887.

[182] P. Tsilifis, I. Bilionis, I. Katsounaros, and N. Zabaras. "Computationally Efficient Variational Approximations for Bayesian Inverse Problems". In: *Journal of Verification, Validation and Uncertainty Quantification* 1.3, 031004 (2016), pp. 1–13. DOI: 10.1115/1.4034102.

[183] J. A. Hartigan. *Bayes Theory*. Springer Series in Statistics. New York: Springer-Verlag, 1983. DOI: 10.1007/978-1-4613-8242-3.

[184] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2000. DOI: 10.1002/9780470316870.

[185] J. K. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis: Theory and Methods*. Springer Texts in Statistics. New York: Springer, 2006. DOI: 10.1007/978-0-387-35433-0.

[186] N. Bleistein and R. A. Handelsman. *Asymptotic Expansions of Integrals*. Mineola, New York, USA: Dover Publications, Inc., 1986.

[187] R. Wong. *Asymptotic Approximations of Integrals*. Computer Science and Scientific Computing. San Diego, California, USA: Academic Press, Inc., 1989. DOI: 10.1016/c2013-0-07651-7.

[188] W. Paulsen. *Asymptotic Analysis and Perturbation Theory*. Boca Raton, Florida, USA: CRC Press, 2014. DOI: 10.1201/b15165.

[189] N. M. Temme. *Asymptotic Methods for Integrals*. Series in Analysis 6. Singapore: World Scientific Publishing Co. Pte. Ltd., 2015. DOI: 10.1142/9195.

[190] L. Tierney and J. B. Kadane. "Accurate Approximations for Posterior Moments and Marginal Densities". In: *Journal of the American Statistical Association* 81.393 (1986), pp. 82–86. DOI: 10.1080/01621459.1986.10478240.

[191] L. Tierney, R. E. Kass, and J. B. Kadane. "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions". In: *Journal of the American Statistical Association* 84.407 (1989), pp. 710–716. DOI: 10.1080/01621459.1989.10478824.

[192] L. Tierney, R. E. Kass, and J. B. Kadane. "Approximate marginal densities of nonlinear functions". In: *Biometrika* 76.3 (1989), pp. 425–433. DOI: 10.1093/biomet/76.3.425.

[193] C. Papadimitriou, J. L. Beck, and L. S. Katafygiotis. "Asymptotic Expansions for Reliability and Moments of Uncertain Systems". In: *Journal of Engineering Mechanics* 123.12 (1997), pp. 1219–1229. DOI: 10.1061/(ASCE)0733-9399(1997)123:12(1219).

[194] S. K. Au, C. Papadimitriou, and J. L. Beck. "Reliability of uncertain dynamical systems with multiple design points". In: *Structural Safety* 21.2 (1999), pp. 113–133. DOI: 10.1016/S0167-4730(99)00009-0.

[195] D. C. Polidori, J. L. Beck, and C. Papadimitriou. "New Approximations for Reliability Integrals". In: *Journal of Engineering Mechanics* 125.4 (1999), pp. 466–475. DOI: 10.1061/(ASCE)0733-9399(1999)125:4(466).

[196] J. He, X. Guan, and R. Jha. "Improve the Accuracy of Asymptotic Approximation in Reliability Problems Involving Multimodal Distributions". In: *IEEE Transactions on Reliability* 65.4 (2016), pp. 1724–1736. DOI: 10.1109/TR.2016.2604121.

[197] B. Bornkamp. "Approximating Probability Densities by Iterated Laplace Approximations". In: *Journal of Computational and Graphical Statistics* 20.3 (2011), pp. 656–669. DOI: 10.1198/jcgs.2011.10099.

Part II

Published papers

Chapter 4

Multilevel uncertainty quantification in Bayesian inverse problems

Original publication

J. B. Nagel and B. Sudret. “A unified framework for multilevel uncertainty quantification in Bayesian inverse problems”. In: *Probabilistic Engineering Mechanics* 43 (2016), pp. 68–84. DOI: 10.1016/j.probengmech.2015.09.007

Abstract

In this paper a unified probabilistic framework for solving inverse problems in the presence of epistemic and aleatory uncertainty is presented. The aim is to establish a flexible theory that facilitates Bayesian data analysis in experimental scenarios as they are commonly met in engineering practice. Problems are addressed where learning about unobservable inputs of a forward model, e.g. reducing the epistemic uncertainty of fixed yet unknown parameters and/or quantifying the aleatory uncertainty of variable inputs, is based on processing response measurements. Approaches to Bayesian inversion, hierarchical modeling and uncertainty quantification are combined into a generic framework that eventually allows to interpret and accomplish this task as multilevel model calibration. A joint problem formulation, where quantities that are not of particular interest are marginalized out from a joint posterior distribution, or an intrinsically marginal formulation, which is based on an integrated likelihood function, can be chosen according to the inferential objective and computational convenience. Fully Bayesian probabilistic inversion, i.e. the inference the variability of unobservable model inputs across a number of experiments, is derived as a special case of multilevel inversion. Borrowing strength, i.e. the optimal estimation of experiment-specific unknown forward model inputs, is introduced as a means for combining information in inverse problems. Two related statistical models for situations involving finite or zero model/measurement error are devised. Multilevel-specific obstacles to Bayesian posterior computation via Markov chain Monte Carlo are discussed. The inferential machinery of Bayesian multilevel model calibration and its underlying flow of information are studied on the basis of a system from the domain of civil engineering. A population of identically manufactured structural elements serves as an exemplary system for examining different experimental settings from the standpoint of uncertainty quantification and reduction. In a series of tests the material variability throughout the ensemble of specimens, the entirety of specimen-specific material properties and the measurement error level are inferred under various uncertainties in the problem setup.

4.1 Introduction

Main characteristics and challenges of inverse problems in engineering sciences subsume the following issues. Firstly, the ever-growing complexity of physical modeling increases the computational expense of deterministic forward simulations. Secondly, uncertainty is omnipresent and calls for an adequate mathematical formalism of representation and management. Thirdly, since data are commonly scarce or prohibitively expensive to acquire, the available information has to be carefully handled. An abstract inverse problem statement thus reads as follows. By analyzing a limited amount of data the endeavor is to optimally learn about unknown forward model inputs that are subject to epistemic uncertainty and aleatory variability. This includes deducing fixed albeit unknown forward model parameters as well as hyperparameters that determine the distribution of variable model inputs. Such a universal formulation describes a class of inverse problems that has hardly been satisfactorily

solved yet. Our goal is therefore to develop a rigorous and extensive framework for formulating and solving such inverse problems in support of data analysis for engineering systems. The focus of this research is on experimental situations as they are typically encountered in this field. We emphasize aspects of uncertainty quantification and information accumulation. In order to establish a sound conceptional and computational basis for solving those problems one has to complement ideas and techniques that have been developed in different academic disciplines and scientific communities so far. This involves inverse modeling, Bayesian statistics and uncertainty quantification. In the following we will shortly survey relevant theories and practices.

In the first place we rely on the Bayesian approach to *classical inverse problems* [1, 2]. When a physical theory or a computational solver relates physical parameters to measurable quantities, i.e. the *forward model*, classical inversion is the process of reasoning or inferring unknown yet physically fixed model parameters from recorded data [3, 4]. Bayesian inference establishes a convenient probabilistic framework to accomplish this conventional type of parameter estimation and data assimilation. At least since the advent of the personal computer it is nowadays widely used in engineering applications [5, 6]. The stochastic paradigm provides a natural mechanism for the regularization of ill-posed problems, however, it requires the specification of a prior and a noise model. *Hierarchical inversion* is an extension of the classical framework that allows to set parameters of the prior and the noise model in a data-informed manner [7, 8]. While epistemic uncertainty is naturally incorporated, a shortcoming of these types of parameter estimation is that they do not account for aleatory variability.

In the second place *hierarchical statistical models* serve as the main tool for the analysis of complex systems. Those are systems that are hierarchically organized at multiple nested layers. Prominent instances include *random* and *mixed effects models* [9]. Historically those models were developed in social and biological sciences e.g. for purposes of educational research [10, 11] and pharmacokinetics/dynamics [12, 13]. Some recent reviews about the methods that were developed in these fields can be found in [14, 15]. Hierarchical modeling can be viewed from a more frequentist [16, 17] or a more Bayesian perspective [18, 19]. At the present day it is mature area of research that establishes sort of an overarching theme in modern multidisciplinary statistics. Dedicated chapters can be found in numerous standard references for Bayesian modeling and inference [20, 21]. A general observation is that hierarchical models may be complex in their probabilistic architecture whereas only little forward modeling takes place.

In the third place we respect the uncertainty taxonomy that is prevalent in risk assessment and decision making. According to this classification one distinguishes between epistemic and aleatory uncertainty [22, 23]. On one side, *epistemic uncertainty* refers to the ignorance or lack of knowledge of the observer and analyst. By taking further evidence this type of uncertainty is reducible in principle. On the contrary, *aleatory uncertainty* or *variability* refers to a trait of the system under consideration. It is a structural randomness of irreducible character. Uncertainties can be accounted for in distinct mathematical frameworks and especially the representation of ignorance is the subject matter of ongoing debates [24, 25]. Graphical statistical models such as *Bayesian probability networks* establish a powerful and widespread tool of uncertainty characterization [26, 27]. In risk-based decision making Bayesian belief networks have been adopted for their strength and flexibility in uncertainty modeling [28, 29] and their elegant mechanisms of information aggregation [30, 31].

In the fourth place *probabilistic inverse problems* constitute a challenging class of inverse problems that is of theoretical and practical relevance alike. While classical inversion is concerned with estimating uncertain yet physically fixed parameters in a series of experiments, i.e. identifying an epistemically uncertain quantity, probabilistic inversion deals with inferring the distribution of such forward model inputs that vary throughout the experiments, i.e. quantifying their aleatory variability. Previously established approaches to this interesting type of problems with *latent/hidden variable* structure subsume various approximate solutions. A frequentist technique that is premised on the simulation of an explicitly marginalized likelihood is proposed in [32]. There are also attempts to compute approximate solutions based on variants of the expectation-maximization algorithm within a linearized Gaussian frame [33] or with the aid of Kriging surrogates [34]. A methodological review of this school of probabilistic inversion is found in [35]. These methods are only partly Bayesian and suffer from the deficiency of providing mere point estimates.

The potential of hierarchical models as instruments of statistical modeling and uncertainty quantification have barely been acknowledged for the purposes of inversion in a classical sense. Hierarchical and probabilistic inversion are first steps towards preparing the Bayesian framework for the treatment of more realistic experimental scenarios. These approaches do not fully exhaust the inferential machinery of hierarchical models and the probability logic of Bayesian networks, though. In this contribution we thus aim at bridging that gap by developing a coherent Bayesian framework for managing uncertainties in such undertakings. By drawing on the statistical theory of hierarchical models, we cast inversion under parameter uncertainty and variability as *Bayesian multilevel calibration*. This embeds a joint and a marginal problem formulation of Bayesian inference under uncertainty, both of which can be numerically solved with plain vanilla or specialized Markov chain Monte

Carlo methods.

This new formulation of *multilevel inversion* is especially well-adapted to the challenges that engineers are frequently faced with. It naturally allows for sophisticated uncertainty modeling which comprises both epistemic and aleatory uncertainty. The inclusion of the former is straightforward whereas the introduction of the latter is an extension to classical parameter estimation. It also promotes a pervasive "blackbox" point of view on the forward model. While this is inevitable in many complex applications, it is not readily compliant with traditional hierarchical models. Previously established strategies of enhanced uncertainty quantification, e.g. hierarchical and probabilistic inversion, emerge as special cases of the proposed general problem formulation. This also offers the opportunity to cope with probabilistic inversion within a fully Bayesian setting. Beyond these extensions some fundamentally new possibilities are suggested. Based on the probabilistic calculus of multilevel models, we develop a novel formulation of multilevel inversion in the zero-noise and "perfect" data limit. The statistical effect of "borrowing strength" or "optimal combination of information" is transferred and applied to inverse problems.

The article is organized as follows. In Section 4.2 we will elaborate a general Bayesian framework for the treatment of uncertainty and variability in inverse problems. This is followed by a discussion about Bayesian inference in the context of multilevel inversion in Section 4.3. Thereafter Section 4.4 will provide an extension of the framework that will allow for handling "perfect" data. Probabilistic inversion and borrowing strength will be placed in context in Sections 4.5 and 4.6, respectively. Dedicated Bayesian computations based on Markov chain Monte Carlo are reviewed in Section 4.7. Lastly in Section 4.8 we will conduct a selection of numerical case studies, where by considering various experimental situations and uncertainty setups the very potential and the computational challenges of the devised modeling paradigm will become transparent.

4.2 Bayesian multilevel modeling

Due to the lack of a unified terminology, we define a *hierarchical* or *multilevel model* as "an overall system model that is hierarchically composed of deterministic and stochastic submodels". Important types of submodels comprise physical models of the deterministic system components (Section 4.2.1), prior descriptions of parameter uncertainty and variability (Section 4.2.2) and residual representations of forward model prediction errors (Section 4.2.3). From these submodels we will assemble a generic Bayesian multilevel model (Section 4.2.4). This will represent the overall system under consideration including its deterministic and probabilistic aspects.

4.2.1 Forward model: Deterministic subsystem

A so-called *forward model* is a mathematical representation of the physical system or phenomenon under investigation. More formally the forward model is a function

$$\begin{aligned} \mathcal{M}\colon \mathcal{D}_{\boldsymbol{m}} \times \mathcal{D}_{\boldsymbol{x}} \times \mathcal{D}_{\boldsymbol{\zeta}} \times \mathcal{D}_{\boldsymbol{d}} &\to \mathcal{D}_{\tilde{\boldsymbol{y}}} \\ (\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) &\mapsto \tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}), \end{aligned} \tag{4.1}$$

that maps inputs $(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \in \mathcal{D}_{\boldsymbol{m}} \times \mathcal{D}_{\boldsymbol{x}} \times \mathcal{D}_{\boldsymbol{\zeta}} \times \mathcal{D}_{\boldsymbol{d}}$ from its domain to outputs $\tilde{\boldsymbol{y}} \in \mathcal{D}_{\tilde{\boldsymbol{y}}}$ from its codomain. Forward model arguments $(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d})$ constitute physical parameters, while its responses $\tilde{\boldsymbol{y}}$ are predictions of observable quantities.

We distinguish between four different types of forward model inputs. They differ in their (un)certain nature when a number of experiments is carried out. There are fixed albeit unknown model parameters $\boldsymbol{m} \in \mathcal{D}_{\boldsymbol{m}}$ that are subject to epistemic uncertainty, two different types of inputs $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ and $\boldsymbol{\zeta} \in \mathcal{D}_{\boldsymbol{\zeta}}$ that are subject to aleatory variability and well-known experimental conditions $\boldsymbol{d} \in \mathcal{D}_{\boldsymbol{d}}$.

4.2.2 Prior model: Input uncertainty

Forward model inputs $\boldsymbol{d}$ constitute perfectly known conditions that prevail during experimentation. In line with this they are deterministic arguments of the forward model. Experimental conditions may differ throughout the experiments, i.e. each of the experiments $i = 1, \ldots, n$ is conducted being subject to an experiment-specific condition $\boldsymbol{d}_i$.

Proper forward model parameters $\boldsymbol{m}$ are constant throughout the experiments $i = 1, \ldots, n$, yet they have unknown values. In Bayesian fashion the available prior or expert knowledge about the true parameter values is represented as a random variable or vector

$$\boldsymbol{M} \sim \pi_{\boldsymbol{M}}(\boldsymbol{m}). \tag{4.2}$$

The Bayesian prior distribution $\pi_{\boldsymbol{M}}(\boldsymbol{m})$ quantifies a subjective degree of plausibility or belief about the true parameter values $\boldsymbol{m}$. This is the Bayesian account for *epistemic uncertainty*. The uncertainty is reducible in the sense that Bayesian data analysis gives rise to a posterior probability model.

Forward model inputs $\boldsymbol{\zeta}$ are subject to a form of variability that is well-known, e.g. it could be ascertained in previous experiments or due to prior considerations. Rather than being constant throughout the experiments $i = 1, \dots, n$, these variable inputs take on experiment-specific realizations $\boldsymbol{\zeta}_i$, all of which are unknown. The corresponding Bayesian prior representation is as mutually independent random variables

$$\boldsymbol{Z}_i \sim f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i}), \text{ for } i = 1, \dots, n. \tag{4.3}$$

Distributions $f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i})$ specify prior knowledge about the experiment-specific unknowns that is of structural quality. They are prescribed by well-known hyperparameters $\boldsymbol{\theta}_{\boldsymbol{Z}_i} \in \mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{Z}}}$, e.g. shape, scale and dependency parameters, that possibly differ across the experiments. Due to stochastic independence, the appropriate joint Bayesian prior model follows as

$$(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n) \sim \prod_{i=1}^{n} f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i}). \tag{4.4}$$

This is a Bayesian conception of *aleatory variability*, i.e. an uncertainty that is of structural nature. Hereinafter this probability model will also be referred to as *prescribed uncertainty*. It is irreducible in the sense that by Bayesian data analysis of the experiments $i = 1, \dots, n$ "past" realizations $\boldsymbol{\zeta}_i$ can be inferred in principle, whereas the knowledge about "future" realizations $\boldsymbol{\zeta}_{i'}$ in further experiments $i' = n+1, \dots, n+n'$ cannot be improved. "Future" realizations still feature a structural uncertainty $\boldsymbol{Z}_{i'} \sim f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_{i'}; \boldsymbol{\theta}_{\boldsymbol{Z}_{i'}})$ that is prescribed by hyperparameters $\boldsymbol{\theta}_{\boldsymbol{Z}_{i'}} \in \mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{Z}}}$.

Another Bayesian notion of a similar type allows to account for forward model inputs $\boldsymbol{x}$ that are subject to a sort of variability which itself is unknown. For $1 = 1, \dots, n$ these variables take on experiment-specific realizations $\boldsymbol{x}_i$, neither of which are known. Bayesian prior modeling is build upon conditionally independent random variables

$$(\boldsymbol{X}_i | \boldsymbol{\Theta}_{\boldsymbol{X}} = \boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}), \text{ for } i = 1, \dots, n. \tag{4.5}$$

The conditional probability distribution $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})$ represents a structural kind of prior knowledge about the experiment-specific unknowns. Its determining hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}} \in \mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{X}}}$, e.g. location, dispersion and correlation parameters, themselves are fixed yet unknown. Hence these hyperparameters are priorly modeled as a random vector

$$\boldsymbol{\Theta}_{\boldsymbol{X}} \sim \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.6}$$

The Bayesian prior distribution $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ constitutes the subjective prior belief or available prior knowledge about the true hyperparameter values. In the statistical literature hyperprior elicitation is exhaustively discussed especially for variance hyperparameters [36–38]. Consequently the joint distribution of the unknowns of this prior model is given as

$$(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n, \boldsymbol{\Theta}_{\boldsymbol{X}}) \sim \left(\prod_{i=1}^{n} f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \right) \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.7}$$

The joint prior distribution of experiment-specific realizations follows by marginalizing Eq. (4.7) over the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$. Then one has

$$(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \sim \int_{\mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{X}}}} \left(\prod_{i=1}^{n} f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \right) \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}) \, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}. \tag{4.8}$$

This is a form of *exchangeability* [39, 40] that realizes some "similarity" of the intermediate variables, i.e. the joint distribution of the sequence $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ equals the one of $(\boldsymbol{X}_{\tau(1)}, \dots, \boldsymbol{X}_{\tau(n)})$ for any index permutation $\tau \colon \{1, \dots, n\} \to \{1, \dots, n\}$. In the present form Eq. (4.8), exchangeability establishes another Bayesian approach to aleatory variability. Unlike the prescribed uncertainty in Eq. (4.4), this form of uncertainty is partially reducible in the sense that the "fuzziness" inherent in Eq. (4.8) can be reduced by learning about $\boldsymbol{\theta}_{\boldsymbol{X}}$ in "past" experiments i. "Past" realizations $\boldsymbol{x}_i$ can also be inferred, however, even if the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ would be known, the realizations $\boldsymbol{x}_{i'}$ of "future" experiments i' would still carry the structural prior uncertainty $\boldsymbol{X}_{i'} \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_{i'} | \boldsymbol{\theta}_{\boldsymbol{X}})$.

In short, on the one hand we have *parametric priors* $\pi_{\boldsymbol{M}}(\boldsymbol{m})$ and $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ that in Eqs. (4.2) and (4.6) embody knowledge about global unknowns $\boldsymbol{m}$ and $\boldsymbol{\theta}_{\boldsymbol{X}}$. On the other hand we have *structural priors* $f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i})$ and $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})$ that encapsulate structural prior knowledge about the problem, and that for $i = 1, \dots, n$ establish the prior model of experiment-specific unknowns $\boldsymbol{x}_i$ and $\boldsymbol{\zeta}_i$ through Eqs. (4.4) and (4.8).

4.2.3 Residual model: Output imperfection

Besides a representation of forward model input uncertainty and variability, an integral constituent of statistical approaches to inversion is a *residual representation* of forward model output discrepancy or imperfection. Due to measurement errors, numerical approximations and general inadequacies, even if all inputs $(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ were perfectly known, predictions $\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ are expected to deviate from real observations $\boldsymbol{y}_i$. These imperfections can be accounted for by a *statistical data model*

$$\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i) + \boldsymbol{\varepsilon}_i, \;\; \text{for } i = 1, \dots, n, \tag{4.9}$$

where residual terms $\boldsymbol{\varepsilon}_i \in \mathcal{D}_{\boldsymbol{\varepsilon}}$ are assumed to be realizations of random variables $\boldsymbol{E}_i \sim f_{\boldsymbol{E}}(\boldsymbol{\varepsilon}_i; \boldsymbol{\Sigma}_i)$. Commonly one employs normal distributions $f_{\boldsymbol{E}}(\boldsymbol{\varepsilon}_i; \boldsymbol{\Sigma}_i) = \mathcal{N}(\boldsymbol{\varepsilon}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_i)$ with mean $\boldsymbol{0}$ and possibly experiment-specific, symmetric and positive-semidefinite covariance matrices $\boldsymbol{\Sigma}_i$. Consequently, through a change of variables whose Jacobian determinant equals one, observations are viewed as outcomes $\boldsymbol{y}_i$ of random variables

$$(\boldsymbol{Y}_i | \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i = \boldsymbol{\zeta}_i) \sim f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big), \;\; \text{for } i = 1, \dots, n. \tag{4.10}$$

For given values of the direct forward model inputs $(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$, data are viewed as random variables $(\boldsymbol{Y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i)$ with conditional distributions $f(\boldsymbol{y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i) = f_{\boldsymbol{E}}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i)$. Note that $f(\boldsymbol{y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i) = f(\boldsymbol{y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{\theta}_{\boldsymbol{X}})$ is independent of $\boldsymbol{\theta}_{\boldsymbol{X}}$.

The specification of the residual model, i.e. quantifying the parameters of $\boldsymbol{\Sigma}_i$, is an essential part of calibrating the forward model and the experimental apparatus. In many experimental situations a model of the *prediction error* is not known a priori, though. Nevertheless, the structure of the prediction error model can be selected [41] and its parameters can be introduced as unknown hyperparameters that undergo calibration [42]. This also includes systematic forward model deviations [43, 44]. Moreover one could treat the form of the forward model $\mathcal{M}$ itself as uncertain/random [45, 46] and select the most plausible class via Bayesian model selection [47, 48]. By adding another layer of uncertainty on top of the outlined setup and at a higher associated cost, the aforementioned principles of assessing structural and parametric forward model uncertainty can be readily applied in multilevel models [49].

Based on random variable transformations, in Section 4.4 we will extend the framework by a model for analyzing "perfect" observations $\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ in the zero-noise limit $|\boldsymbol{\varepsilon}_i| \to 0$. This mathematical formulation will explain the variability in the data exclusively by a Bayesian prior model of input variability as outlined in the preceding Section 4.2.2.

4.2.4 Multilevel model: Overall system

We start from the premise that if not denoted or stated otherwise, random vectors and variables are (conditionally) independent, e.g. the global forward model parameters $\boldsymbol{M}$ and the hyperparameters $\boldsymbol{\Theta}_{\boldsymbol{X}}$ are understood to be priorly independent. Thus $\pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \pi_{\boldsymbol{M}}(\boldsymbol{m})\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ applies for their joint prior distribution. Note that this is not a necessity of the formulation, though. Moreover, we strictly reserve conditional notation for the stochastic dependency of random variables on outcomes of other random variables, e.g. the aleatory variables $(\boldsymbol{X}_i | \boldsymbol{\theta}_{\boldsymbol{X}})$ are conditionally dependent on realizations $\boldsymbol{\Theta}_{\boldsymbol{X}} = \boldsymbol{\theta}_{\boldsymbol{X}}$. The stochastic variables $(\boldsymbol{Y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i)$ are conditioned on random outcomes $\boldsymbol{M} = \boldsymbol{m}$, $\boldsymbol{X}_i = \boldsymbol{x}_i$ and $\boldsymbol{Z}_i = \boldsymbol{\zeta}_i$, nonetheless they depend on deterministic quantities $\boldsymbol{d}_i$ and $\boldsymbol{\Sigma}_i$, too. Similarly the aleatory variables $\boldsymbol{Z}_i$ are dependent on $\boldsymbol{\theta}_{\boldsymbol{Z}_i}$ in a way that is not explicitly indicated. In order to keep track of all stochastic and deterministic relations the index i serves as a bookkeeping mark.

Deterministic aspects of the system are covered by the forward model Eq. (4.1). Parametric priors in Eqs. (4.2) and (4.6) and structural priors in Eqs. (4.3) and (4.5) represent input uncertainty and variability. The model Eq. (4.10) condenses basic assumptions regarding the prediction error. Altogether those submodels are combined into a greater model of the whole system. The overall probability model is summarized as

$$(\boldsymbol{Y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i) \sim f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big), \tag{4.11a}$$

$$\boldsymbol{M} \sim \pi_{\boldsymbol{M}}(\boldsymbol{m}), \tag{4.11b}$$

$$\boldsymbol{Z}_i \sim f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i}), \tag{4.11c}$$

$$(\boldsymbol{X}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}), \tag{4.11d}$$

$$\boldsymbol{\Theta}_{\boldsymbol{X}} \sim \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.11e}$$

Adopting a subjectivist viewpoint, this complex probability model Eq. (4.11) formalizes degrees of belief of how the data have been realized in the experiments $i = 1, \dots, n$. According to our previous definition it is a generic Bayesian multilevel model. An intuitive representation of this multilevel model is provided by a directed acyclic graph (DAG) [26, 27] such as shown in Fig. 4.1.

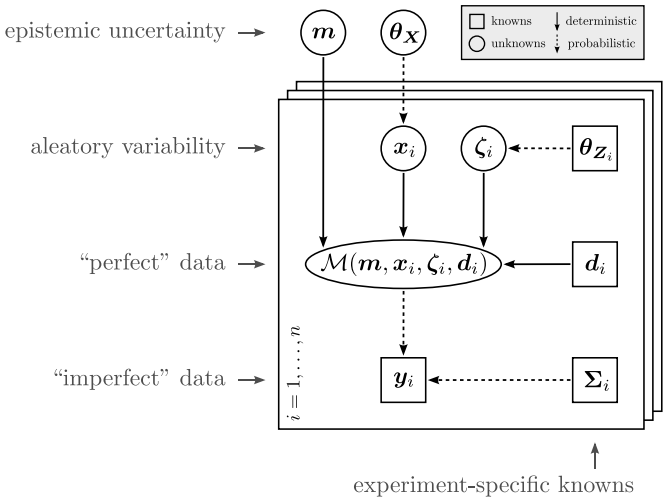


Figure 4.1: DAG of the generic multilevel model. Vertices symbolize known (□) or unknown (○) quantities, while directed edges represent their deterministic (⟶) or probabilistic (⇢) relations. Global parameters $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ are subject to epistemic uncertainty, whereas experiment-specific realizations $(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$ are subject to aleatory variability. Known quantities comprise the data $\langle \boldsymbol{y}_i \rangle$ just as well as experiment-specific knowns $(\langle \boldsymbol{\theta}_{\boldsymbol{Z}_i} \rangle, \langle \boldsymbol{d}_i \rangle, \langle \boldsymbol{\Sigma}_i \rangle)$ located at different levels of the hierarchy.

4.3 Inference in multilevel models

We will now discuss statistical inference. In particular we will demonstrate how conditioning on the observables and marginalization out nuisance are elegant inferential tools of Bayesian multilevel inversion. A pivotal joint problem formulation will be devised. Afterwards an intrinsically marginal problem variant will be presented in Section 4.3.1.

In the following $\langle \boldsymbol{q}_i \rangle$ denotes a sequence $\langle \boldsymbol{q}_i \rangle_{1 \leq i \leq n} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_n)$. Summarizing the available parametric and structural prior knowledge in Eqs. (4.11b) to (4.11e), the *joint prior* of the entirety of unknowns $(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}})$ factorizes as

$$\pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}}\big) = \left(\prod_{i=1}^{n} f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \right) \left(\prod_{i=1}^{n} f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i}) \right) \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})\, \pi_{\boldsymbol{M}}(\boldsymbol{m}). \tag{4.12}$$

This prior depends only on the collection of experiment-specific hyperparameters $\langle \boldsymbol{\theta}_{\boldsymbol{Z}_i} \rangle$. With the model of single observations in Eq. (4.11a) one can formulate a conditional distribution for the total data $\langle \boldsymbol{y}_i \rangle$. For given values of the unknowns $(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$ this yields the product $f(\langle \boldsymbol{y}_i \rangle | \boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle) = \prod_{i=1}^{n} f_{\boldsymbol{E}}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i)$. It depends on experiment-specific knowns $(\langle \boldsymbol{d}_i \rangle, \langle \boldsymbol{\Sigma}_i \rangle)$.

With that said, one can derive the *joint posterior* of the totality of unknowns $(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}})$ by conditioning on the acquired data $\langle \boldsymbol{y}_i \rangle$. By virtue of Bayes' theorem one obtains

$$\pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle\big) = \frac{1}{C} \left(\prod_{i=1}^{n} f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big) \right) \pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}}\big). \tag{4.13}$$

This posterior Eq. (4.13) is implicitly dependent on experiment-specific knowns $(\langle \boldsymbol{\theta}_{\boldsymbol{Z}_i} \rangle, \langle \boldsymbol{d}_i \rangle, \langle \boldsymbol{\Sigma}_i \rangle)$. It is the central object in Bayesian multilevel model calibration.

The model evidence C is the total probability of the realized data $\langle \boldsymbol{y}_i \rangle$, given the underlying multilevel model. When introducing the notation $\mathrm{d}\langle \boldsymbol{q}_i \rangle = \mathrm{d}\boldsymbol{q}_1\, \mathrm{d}\boldsymbol{q}_2 \dots \mathrm{d}\boldsymbol{q}_n$ one can write this as

$$C = \int_{\mathcal{D}_{\boldsymbol{m}}} \int_{\mathcal{D}_{\boldsymbol{x}}^n} \int_{\mathcal{D}_{\boldsymbol{\zeta}}^n} \int_{\mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{X}}}} \left(\prod_{i=1}^{n} f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big) \right) \pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}}\big)\, \mathrm{d}\boldsymbol{m}\, \mathrm{d}\langle \boldsymbol{x}_i \rangle\, \mathrm{d}\langle \boldsymbol{\zeta}_i \rangle\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}. \tag{4.14}$$

For the Bayesian computations that will be reviewed in Section 4.7, the factor of proportionality C does not have to be computed explicitly. For that reason it will be occasionally omitted from now on.

One may define a likelihood in order to write the joint posterior Eq. (4.13) in the familiar textbook-form $\pi(\text{unknowns}|\text{data}) \propto \mathcal{L}(\text{unknowns}; \text{data})\, \pi(\text{unknowns})$. Regarded as a function of the unknowns $(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$,

the *joint likelihood* evaluates the densities in Eq. (4.10) for the collected data $\langle \boldsymbol{y}_i \rangle$ by

$$\mathcal{L}\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle; \langle \boldsymbol{y}_i \rangle\big) = \prod_{i=1}^{n} f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big). \tag{4.15}$$

Apart from its functional arguments and the data it also depends on the total number of experiment-specific knowns $(\langle \boldsymbol{d}_i \rangle, \langle \boldsymbol{\Sigma}_i \rangle)$. It does not depend on $\boldsymbol{\theta}_{\boldsymbol{X}}$, though.

Subsequent to formulating the joint posterior Eq. (4.13) the marginal of the *quantities of interest* (QoI) is obtained by integrating out *nuisance* [50, 51]. For instance, given that $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ are declared QoI and the latent variables $(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$ are considered nuisance, the correspondingly marginalized posterior becomes

$$\pi\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}} \,|\, \langle \boldsymbol{y}_i \rangle\big) = \int_{\mathcal{D}_{\boldsymbol{x}}^n} \int_{\mathcal{D}_{\boldsymbol{\zeta}}^n} \pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} \,|\, \langle \boldsymbol{y}_i \rangle\big) \,\mathrm{d}\langle \boldsymbol{x}_i \rangle \,\mathrm{d}\langle \boldsymbol{\zeta}_i \rangle. \tag{4.16}$$

Similarly, provided that hidden variables $(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$ are proclaimed QoI and $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ are deemed nuisance parameters, appropriately marginalizing the posterior distribution gives

$$\pi\big(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle \,|\, \langle \boldsymbol{y}_i \rangle\big) = \int_{\mathcal{D}_{\boldsymbol{m}}} \int_{\mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{X}}}} \pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} \,|\, \langle \boldsymbol{y}_i \rangle\big) \,\mathrm{d}\boldsymbol{m} \,\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}. \tag{4.17}$$

4.3.1 Marginalized formulation

A common scenario is that inferential interest focuses on the global parameters $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$. In this particular case, instead of marginalizing the joint posterior distribution Eq. (4.13) as in Eq. (4.16), based on an integrated likelihood function one can formulate an inherently marginal problem [52–54]. One therefore constructs a marginalized observation model

$$(\boldsymbol{Y}_i \,|\, \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) \sim f(\boldsymbol{y}_i \,|\, \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}), \quad \text{for } i = 1, \ldots, n, \tag{4.18a}$$

$$(\boldsymbol{M}, \boldsymbol{\Theta}_{\boldsymbol{X}}) \sim \pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \pi_{\boldsymbol{M}}(\boldsymbol{m})\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.18b}$$

The marginalized model consists of the prior distribution Eq. (4.18b) of the QoI $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ and the probability model Eq. (4.18a) of the observations $\boldsymbol{y}_i$. By integrating out the aleatory variables $(\boldsymbol{x}_i, \boldsymbol{\zeta}_i)$ in the following way, one can obtain the marginal distributions of the observations

$$f(\boldsymbol{y}_i \,|\, \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \int_{\mathcal{D}_{\boldsymbol{x}}} \int_{\mathcal{D}_{\boldsymbol{\zeta}}} f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big)\, f_{\boldsymbol{X}\,|\,\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}_{\boldsymbol{X}})\, f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i}) \,\mathrm{d}\boldsymbol{x}_i \,\mathrm{d}\boldsymbol{\zeta}_i. \tag{4.19}$$

These distributions are conditional on $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ and dependent on $(\boldsymbol{\theta}_{\boldsymbol{Z}_i}, \boldsymbol{d}_i, \boldsymbol{\Sigma}_i)$. Following this, one can easily formulate an *integrated* or *marginalized likelihood*. Evaluated for the actual data $\langle \boldsymbol{y}_i \rangle$ and seen as a function of the QoI $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ this version of the likelihood reads as

$$\mathcal{L}\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \boldsymbol{y}_i \rangle\big) = f\big(\langle \boldsymbol{y}_i \rangle \,|\, \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}\big) = \prod_{i=1}^{n} f(\boldsymbol{y}_i \,|\, \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.20}$$

It is the likelihood function corresponding to the case of eliminating all intermediate unobservables $(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$ with Eq. (4.19) on the likelihood rather than on the posterior level. Note that frequentist inference of $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ could be based on this integrated likelihood formulation. Fully Bayesian inference, however, proceeds by formulating the corresponding posterior distribution. With the prior Eq. (4.18b) and the likelihood Eq. (4.20), the posterior is obtained on grounds of Bayes' law

$$\pi\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}} \,|\, \langle \boldsymbol{y}_i \rangle\big) = \frac{1}{C}\, \mathcal{L}\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \boldsymbol{y}_i \rangle\big)\, \pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.21}$$

One can easily derive that the normalizing constant C equals Eq. (4.14) and show that the posteriors Eqs. (4.16) and (4.21) are identical. This means that, as far as the inference of $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ is concerned, the two problem formulations Eqs. (4.11) and (4.18) are equivalent. Those problem formulations pose different numerical obstacles, though. In Section 4.7 we will discuss Bayesian computations and their multilevel-related issues.

4.3.1.1 Monte Carlo integration

In Eq. (4.20) the marginalized likelihood $\mathcal{L}(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \boldsymbol{y}_i \rangle)$ is a product of integrals $f(\boldsymbol{y}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$. Most often it is not possible to perform the marginalization in Eq. (4.19) analytically. Still it can be approximately computed through deterministic or stochastic schemes of numerical integration.

The density $f(\boldsymbol{y}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ can be evaluated for arbitrary arguments $\boldsymbol{y}_i$ and for fixed values $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$. A simple numerical means to that end rests upon stochastic integration via the Monte Carlo (MC) method

$$\hat{f}(\boldsymbol{y}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \frac{1}{K} \sum_{k=1}^{K} f_{\boldsymbol{E}} \left(\boldsymbol{y}_i - \tilde{\boldsymbol{v}}_i^{(k)}; \boldsymbol{\Sigma}_i \right), \text{ with } \left\{ \begin{array}{l} \boldsymbol{x}_i^{(k)} \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i^{(k)}|\boldsymbol{\theta}_{\boldsymbol{X}}), \\ \boldsymbol{\zeta}_i^{(k)} \sim f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i^{(k)}; \boldsymbol{\theta}_{\boldsymbol{Z}_i}), \\ \tilde{\boldsymbol{v}}_i^{(k)} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \boldsymbol{d}_i) \end{array} \right\} \text{ for } k = 1, \dots, K. \tag{4.22}$$

For $k = 1, \dots, K$ forward model inputs $\boldsymbol{x}_i^{(k)}$ and $\boldsymbol{\zeta}_i^{(k)}$ are independently sampled from their population distributions $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i^{(k)}|\boldsymbol{\theta}_{\boldsymbol{X}})$ and $f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i^{(k)}; \boldsymbol{\theta}_{\boldsymbol{Z}_i})$, respectively. In turn responses $\tilde{\boldsymbol{v}}_i^{(k)} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \boldsymbol{d}_i)$ are computed accordingly. For evaluating $\mathcal{L}(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \boldsymbol{y}_i \rangle)$ as a function of the unknowns $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$, one has to simulate Eq. (4.22) for the observations $\boldsymbol{y}_i$ that were taken in the experiments $i = 1, \dots, n$. Thus a simple MC-based estimator of the marginalized likelihood is given as

$$\hat{\mathcal{L}}\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \boldsymbol{y}_i \rangle\big) = \prod_{i=1}^{n} \hat{f}(\boldsymbol{y}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.23}$$

The stochastic simulator Eq. (4.23) may be costly and numerically inefficient in terms of the number of runs K of the deterministic model. It should be understood as an instructive proof for the feasibility of computing the marginal posterior Eq. (4.21). In practice more advanced simulators, e.g. based on importance sampling, can be applied in similar fashion [55, 56]. More generally speaking, any method for computing the model evidence in classical Bayesian inference is applicable [57].

4.4 Zero-noise and "perfect" data

In Section 4.2.3 the residual model was introduced as a representation of the discrepancy between model predictions and measurements. This conditional model had equipped the data space $\mathcal{D}_{\tilde{\boldsymbol{y}}}$ with a probability measure. As a consequence, in Eq. (4.11a) observations were regarded as $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i$ with a random outcome $\boldsymbol{\varepsilon}_i$. However, experimental situations may occur where direct access to

$$\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i), \text{ for } i = 1, \dots, n \tag{4.24}$$

is granted, e.g. due to noise-free measurements and a "sufficiently accurate" forward model [58]. The data $\langle \tilde{\boldsymbol{y}}_i \rangle$ is then only explained by uncertainty of the forward model inputs as described in Section 4.2.2, without being subject to prediction errors. Hereafter we will refer this scenario as to involve "perfect" data [59, 60]. A statistical model that is appropriate for "perfect" data can be formulated as

$$(\tilde{\boldsymbol{Y}}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) \sim f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}), \text{ for } i = 1, \dots, n, \tag{4.25a}$$

$$(\boldsymbol{M}, \boldsymbol{\Theta}_{\boldsymbol{X}}) \sim \pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \pi_{\boldsymbol{M}}(\boldsymbol{m})\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.25b}$$

As before, Eq. (4.25b) embodies the available prior knowledge about the unknowns $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$. Conditional random variables in Eq. (4.25a) are constructed by forward uncertainty propagation as follows. The independent input uncertainties $(\boldsymbol{X}_i|\boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})$ and $\boldsymbol{Z}_i \sim f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i})$, that are defined for given $(\boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}_i})$, are propagated through the forward model $\mathcal{M}$, while the inputs $(\boldsymbol{m}, \boldsymbol{d}_i)$ are fixed. The density of the resulting random variables $(\tilde{\boldsymbol{Y}}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \mathcal{M}\big(\boldsymbol{m}, (\boldsymbol{X}_i|\boldsymbol{\theta}_{\boldsymbol{X}}), \boldsymbol{Z}_i, \boldsymbol{d}_i\big)$ at $\tilde{\boldsymbol{y}}_i \in \mathcal{D}_{\tilde{\boldsymbol{y}}}$ is found as

$$f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \int_{\mathcal{D}_{\boldsymbol{x}}} \int_{\mathcal{D}_{\boldsymbol{\zeta}}} \delta\big(\tilde{\boldsymbol{y}}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)\big)\, f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})\, f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i})\, \mathrm{d}\boldsymbol{x}_i\, \mathrm{d}\boldsymbol{\zeta}_i, \tag{4.26}$$

where δ denotes the Dirac delta distribution. This endows the response space $\mathcal{D}_{\tilde{\boldsymbol{y}}}$ with a proper probability model. Inspecting Eqs. (4.19) and (4.26) reveals that the marginal model Eq. (4.18) approaches the "perfect" data model Eq. (4.25) in the zero-noise limit $\|\boldsymbol{\Sigma}_i\| \to 0$. With the distributions $f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$, that are conditioned on

$(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ and dependent on experiment-specific knowns $(\boldsymbol{d}_i, \boldsymbol{\theta}_{\boldsymbol{Z}_i})$, one can formulate the corresponding likelihood function as

$$\mathcal{L}\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \tilde{\boldsymbol{y}}_i \rangle\big) = f\big(\langle \tilde{\boldsymbol{y}}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}\big) = \prod_{i=1}^{n} f(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.27}$$

For given data $\langle \tilde{\boldsymbol{y}}_i \rangle$ it is viewed as a function of the unknowns $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ that also depends on $(\langle \boldsymbol{\theta}_{\boldsymbol{Z}_i} \rangle, \langle \boldsymbol{d}_i \rangle)$. As usual Bayesian data analysis proceeds by conditioning on the data $\langle \tilde{\boldsymbol{y}}_i \rangle$. With the prior Eq. (4.25b) and the likelihood Eq. (4.27) the posterior follows through Bayes' rule

$$\pi\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \tilde{\boldsymbol{y}}_i \rangle\big) = \frac{1}{\tilde{C}} \mathcal{L}\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \tilde{\boldsymbol{y}}_i \rangle\big)\, \pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.28}$$

The factor of proportionality $\tilde{C}$ in the posterior density Eq. (4.28) is given as the marginal probability density of the effectively acquired data $\langle \tilde{\boldsymbol{y}}_i \rangle$. It thus writes $\tilde{C} = \iint \mathcal{L}(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \tilde{\boldsymbol{y}}_i \rangle)\, \pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})\, \mathrm{d}\boldsymbol{m}\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}$.

4.4.1 Kernel density estimation

The likelihood function $\mathcal{L}(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \tilde{\boldsymbol{y}}_i \rangle)$ in Eq. (4.27) is grounded on probability densities $f(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$. Likelihood evaluations therefore require forward uncertainty propagation Eq. (4.26). In the majority of cases this complicated problem can only be approximately solved. A possible approach is to use MC uncertainty propagation in combination with kernel density estimation (KDE) [61].

Let $\mathcal{K}_{\boldsymbol{H}}(\tilde{\boldsymbol{y}}) = |\boldsymbol{H}|^{-1/2} \mathcal{K}(\boldsymbol{H}^{-1/2} \tilde{\boldsymbol{y}})$ be the scaled kernel that is defined by a kernel function $\mathcal{K}$ and the symmetric and positive-definite bandwidth matrix $\boldsymbol{H}$. A KDE of the density $f(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ in Eq. (4.26) as a function of $\tilde{\boldsymbol{y}}_i$ and for fixed values of $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ is given as

$$\hat{f}(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{K}_{\boldsymbol{H}}\left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{v}}_i^{(k)}\right), \text{ with } \left\{ \begin{array}{l} \boldsymbol{x}_i^{(k)} \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i^{(k)} | \boldsymbol{\theta}_{\boldsymbol{X}}), \\ \boldsymbol{\zeta}_i^{(k)} \sim f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i^{(k)}; \boldsymbol{\theta}_{\boldsymbol{Z}_i}), \\ \tilde{\boldsymbol{v}}_i^{(k)} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \boldsymbol{d}_i) \end{array} \right\} \text{ for } k = 1, \ldots, K. \tag{4.29}$$

Analogously to Eq. (4.22), for $k = 1, \ldots, K$ forward model inputs $\boldsymbol{x}_i^{(k)}$ and $\boldsymbol{\zeta}_i^{(k)}$ are randomly drawn from their parent distributions $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i^{(k)} | \boldsymbol{\theta}_{\boldsymbol{X}})$ and $f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i^{(k)}; \boldsymbol{\theta}_{\boldsymbol{Z}_i})$ and responses $\tilde{\boldsymbol{v}}_i^{(k)} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i^{(k)}, \boldsymbol{\zeta}_i^{(k)}, \boldsymbol{d}_i)$ are computed. Subsequently the sample $(\tilde{\boldsymbol{v}}_i^{(1)}, \ldots, \tilde{\boldsymbol{v}}_i^{(K)})$ serves as a proxy for the distribution $f(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$. Estimating $\mathcal{L}(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \tilde{\boldsymbol{y}}_i \rangle)$ is based on evaluating the KDE in Eq. (4.29) for arguments $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ and for the observations $\tilde{\boldsymbol{y}}_i$ corresponding to experiments $i = 1, \ldots, n$. On these grounds, the likelihood function $\mathcal{L}(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \tilde{\boldsymbol{y}}_i \rangle)$ is approximated as

$$\hat{\mathcal{L}}\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}; \langle \tilde{\boldsymbol{y}}_i \rangle\big) = \prod_{i=1}^{n} \hat{f}(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.30}$$

Similarly to Eq. (4.23) this is an expensive statistical estimation program that involves forward uncertainty quantification and tends to require a high number K of calls to the forward code. Further challenges intrinsically related to computing the posterior Eq. (4.28) of the "perfect" data model will be discussed in Section 4.7.

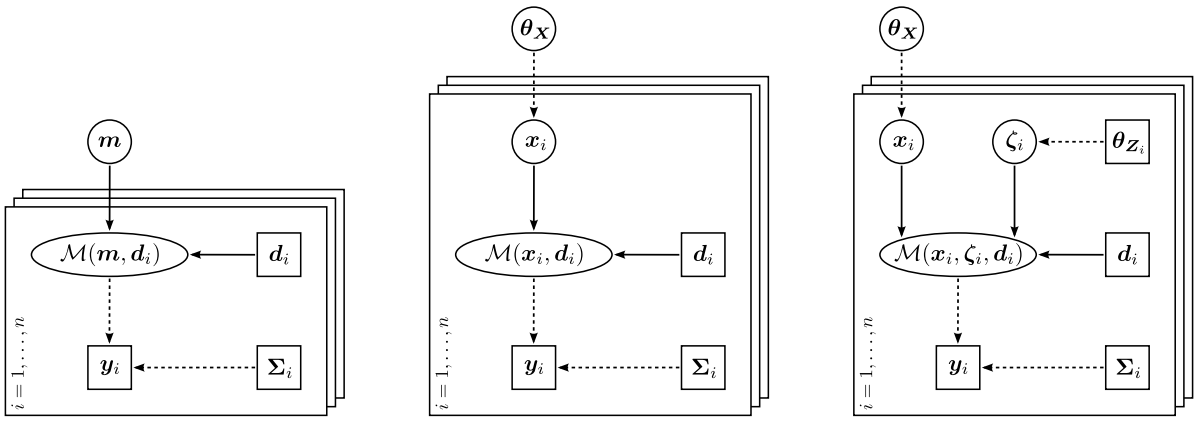
4.5 Probabilistic inversion

The introduced Bayesian multilevel model Eq. (4.11) acts as a toolkit for statistical model building. It forms some kind of superstructure that embeds a variety of stochastic inverse problems as special cases. In this section we will show how different well-known types of inverse problems are obtained by omitting global parameters and/or experiment-specific variables accordingly.

Classical or *simple Bayesian inversion* is concerned with the estimation of fixed yet unknown parameters $\boldsymbol{m}$ of the physical simulator [3, 4]. The related DAG is pictured in Fig. 4.2(a). In this context the term "simple" merely refers to the degree of sophistication of the input uncertainty model. As a matter of fact classical inversion may not be a simple problem at all. It typically calls for a high number of forward solves. The engineering community therefore relies on customized strategies in order to ameliorate the computational burden to Bayesian inference in real-case problems. This includes the employment of polynomial chaos expansions as forward model substitutes [62–64], advanced stochastic simulation techniques [65, 66] and forward model reduction methods [67, 68].

Probabilistic inversion features a more elaborate two-level representation of input uncertainty [69, 70]. Rather than aiming at an unknown constant $\boldsymbol{m}$, inference concentrates on the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ that determine the

variability of $\langle \boldsymbol{x}_i \rangle$ through $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})$. A DAG belonging to probabilistic inversion is depicted in Fig. 4.2(b). Building upon probabilistic inversion one may have variable inputs $\langle \boldsymbol{\zeta}_i \rangle$, the distributions of which $f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i})$ are prescribed by $\langle \boldsymbol{\theta}_{\boldsymbol{Z}_i} \rangle$. Unless experiment-specific realizations of those variables are of inferential interest, they act as additional nuisance parameters impeding the inference of the QoI. The correspondingly extended DAG is provided in Fig. 4.2(c). Of course, more complex modeling scenarios can be envisaged. An application example where inference targets both parameters of the type $\boldsymbol{m}$ and $\boldsymbol{\theta}_{\boldsymbol{X}}$, in the presence of additional nuisance parameters $\langle \boldsymbol{\zeta}_i \rangle$, can be found in [59, 60].



(a) Simple inversion. (b) Probabilistic inversion. (c) Additional nuisance.

Figure 4.2: Various DAGs. Simple inversion, i.e. the estimation of an unknown $\boldsymbol{m}$, is visualized in (a), whereas (b) shows a DAG of probabilistic inversion, i.e. the inference of $\boldsymbol{\theta}_{\boldsymbol{X}}$ that governs the variability of experiment-specific $\boldsymbol{x}_i$. An upgrade of probabilistic inversion, where a prescribed uncertainty has been introduced in nuisance variables $\boldsymbol{\zeta}_i$, is depicted in (c).

The problem that we call probabilistic inversion shall not be confused with the identically named problem of finding an input distribution of a forward model given its output distribution [71, 72]. Commonly engineering applications do not allow to exercise this type of uncertainty backpropagation. The amount and structure of the data being available do not permit to fully specify a response distribution while expert knowledge refers to physical parameters instead.

At this point we have a closer look at probabilistic inversion. It results from removing the forward model inputs $\boldsymbol{m}$ and $\langle \boldsymbol{\zeta}_i \rangle$ from the overall system Eq. (4.11) and from declaring $\boldsymbol{\theta}_{\boldsymbol{X}}$ as QoI and $\langle \boldsymbol{x}_i \rangle$ as nuisance variables. For the sake of completeness we summarize the associated multilevel model as

$$(\boldsymbol{Y}_i|\boldsymbol{x}_i) \sim f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big), \tag{4.31a}$$

$$(\boldsymbol{X}_i|\boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}}), \tag{4.31b}$$

$$\boldsymbol{\Theta}_{\boldsymbol{X}} \sim \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.31c}$$

Joint Bayesian inference is accomplished by conditioning on the realized data $\langle \boldsymbol{y}_i \rangle$. Up to a normalization factor, according to Bayes' law the posterior density is given as

$$\pi\big(\langle \boldsymbol{x}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle\big) \propto \left(\prod_{i=1}^{n} f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big) \right) \left(\prod_{i=1}^{n} f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}}) \right) \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.32}$$

Equivalent to integrating out nuisance $\langle \boldsymbol{x}_i \rangle$ from the joint posterior Eq. (4.32) as in Eq. (4.16), one can base inference of $\boldsymbol{\theta}_{\boldsymbol{X}}$ on an inherently marginal problem formulation [32, 35]. Similar to Eqs. (4.19) and (4.20) the marginalized likelihood function for that case is derived as

$$\mathcal{L}\big(\boldsymbol{\theta}_{\boldsymbol{X}}; \langle \boldsymbol{y}_i \rangle\big) = f\big(\langle \boldsymbol{y}_i \rangle | \boldsymbol{\theta}_{\boldsymbol{X}}\big) = \prod_{i=1}^{n} \int_{\mathcal{D}_{\boldsymbol{x}}} f_{\boldsymbol{E}}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i\big)\, f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})\, \mathrm{d}\boldsymbol{x}_i. \tag{4.33}$$

With the marginalized likelihood function Eq. (4.33) and the marginal prior distribution Eq. (4.31c), the unscaled version of the marginal posterior reduces to

$$\pi\big(\boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle\big) \propto \mathcal{L}(\boldsymbol{\theta}_{\boldsymbol{X}}; \langle \boldsymbol{y}_i \rangle)\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{4.34}$$

Exemplary comparisons of the numerical efficiency for sampling joint posteriors of the form Eq. (4.32) and marginal posteriors of the form Eq. (4.34) are found in [69, 70].

Approximate *two-stage approaches* have been proposed for inferring aleatory parameter variability in inverse problems, e.g. the context of random fields [73–76]. In the first stage n separate inverse problems are solved, i.e. for each experiment $i = 1, \dots, n$ an estimator $\hat{\boldsymbol{x}}_i$ of the realization $\boldsymbol{x}_i$ is computed. As a second step the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ are identified by statistical analysis of the estimates $\langle \hat{\boldsymbol{x}}_i \rangle$. However, two-stage methods suffer from the dependence on a sufficient amount of data available for both of the stages and their tendency to overestimate second-order central moments [14, 15]. Those issues are due to a fundamental inconsistency in treating epistemic and aleatory uncertainty.

Classical inverse problems are sometimes phrased within a hierarchical frame [7, 8]. Formally this is a special case of probabilistic inversion with $n = 1$. The intermediate unknowns $\boldsymbol{x}_1$ are commonly the QoI in this type of *hierarchical inversion*. Their prior $\pi(\boldsymbol{x}_1) = \int f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_1|\boldsymbol{\theta}_{\boldsymbol{X}})\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}$ decomposes into a conditional distribution $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_1|\boldsymbol{\theta}_{\boldsymbol{X}})$ and a marginal one $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$. However, other than in probabilistic inversion, Eq. (4.31b) is not interpreted as aleatory variability. Instead it can be viewed as leaving the prior for $\boldsymbol{x}_1$ incompletely specified [7], i.e. relaxing the assumption of a parametric prior $\pi(\boldsymbol{x}_1; \boldsymbol{\theta}_{\boldsymbol{X}}) = f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_1|\boldsymbol{\theta}_{\boldsymbol{X}})$ for a specific value $\boldsymbol{\theta}_{\boldsymbol{X}}$. Alternatively Eq. (4.32) suggests that the prior hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ can be estimated along with $\boldsymbol{x}_1$. The prior in this case is given as $\pi(\boldsymbol{x}_1, \boldsymbol{\theta}_{\boldsymbol{X}}) = f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_1|\boldsymbol{\theta}_{\boldsymbol{X}})\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$. For solving ill-posed problems this can be seen as an automatic determination of the regularization parameters [8].

4.6 Combination of information

In the preceding Section 4.5 we declared the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ as QoI and latent quantities $\langle \boldsymbol{x}_i \rangle$ as nuisance. When this choice is reversed, i.e. proclaiming $\langle \boldsymbol{x}_i \rangle$ as the QoI and treating $\boldsymbol{\theta}_{\boldsymbol{X}}$ as nuisance, then the Bayesian multilevel model Eq. (4.31) allows for an optimal type of inference [77]. This effect is sometimes referred to as *optimal combination of information* or *borrowing strength*. To our best knowledge, it has been pointed out for the first time in [78]. As we will see, the term "optimal" has to be understood with respect to the total amount of information processed, e.g. the acquired data and the available parametric and structural prior knowledge. Optimal combination of information seems to be largely understudied in inverse problems with missing data structure. By taking the marginal viewpoint of Eq. (4.34), the additional advantages that the joint formulation Eq. (4.32) offers are often overlooked.

Based on the hierarchical model Eq. (4.31), in this section we will show how to "borrow strength" in inverse problems. The optimal inference of a specific $\boldsymbol{x}_{i_0}$ for some $i_0 \in \{1, \dots, n\}$ is demonstrated. We pursue three different estimation programs in order to investigate how inferring $\boldsymbol{x}_{i_0}$ can be accomplished by wholly or only partially utilizing the informational resources. In Section 4.6.1 we will present a simple Bayesian updating approach, in respect to which the principle and mechanism of borrowing strength is emphasized by means of multilevel inference in Section 4.6.3. Beforehand we will devise a sequential filtering approach in Section 4.6.2 that will serve as an illustration of the underlying flow of information.

4.6.1 Simple updating

In this first approach, inference of $\boldsymbol{x}_{i_0}$ will be solely based on the single observation $\boldsymbol{y}_{i_0}$, the informational content of $f_{\boldsymbol{E}}(\boldsymbol{y}_{i_0} - \mathcal{M}(\boldsymbol{x}_{i_0}, \boldsymbol{d}_{i_0}); \boldsymbol{\Sigma}_{i_0})$, the structural prior $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_{i_0}|\boldsymbol{\theta}_{\boldsymbol{X}})$ and the parametric prior $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$. Utilizing the prior information one can formulate a Bayesian prior distribution for $\boldsymbol{x}_{i_0}$. By marginalizing over the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ this reads as

$$\pi(\boldsymbol{x}_{i_0}) = \int_{\mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{X}}}} f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_{i_0}|\boldsymbol{\theta}_{\boldsymbol{X}})\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}. \tag{4.35}$$

This compound distribution represents the uncertainty that $\boldsymbol{x}_{i_0}$ priorly carries. Ensuing from the prior Eq. (4.35), analyzing the piece of data $\boldsymbol{y}_{i_0}$ is accomplished by constructing the corresponding posterior. It is proportional to $\pi(\boldsymbol{x}_{i_0}|\boldsymbol{y}_{i_0}) \propto f_{\boldsymbol{E}}(\boldsymbol{y}_{i_0} - \mathcal{M}(\boldsymbol{x}_{i_0}, \boldsymbol{d}_{i_0}); \boldsymbol{\Sigma}_{i_0})\, \pi(\boldsymbol{x}_{i_0})$. We remark that the approach is formally reminiscent of hierarchical inversion as discussed in Section 4.5.

While the observation $\boldsymbol{y}_{i_0}$ that is directly associated to $\boldsymbol{x}_{i_0}$ has been analyzed, the evidence that $\langle \boldsymbol{y}_{\neq i_0} \rangle$ carry about $\boldsymbol{\theta}_{\boldsymbol{X}}$, and in turn about $\boldsymbol{x}_{i_0}$, has not yet been taken into consideration. Put another way, the hierarchical problem structure has been respected by formulating Eq. (4.35), however, it has only been partially exploited for learning about the QoI $\boldsymbol{x}_{i_0}$.

4.6.2 Sequential filtering

For the second estimation scheme, which will be based on sequential updating, we introduce the simplifying notation $\langle \boldsymbol{q}_{\neq i_0} \rangle = (\boldsymbol{q}_1, \dots, \boldsymbol{q}_{i_0-1}, \boldsymbol{q}_{i_0+1}, \dots, \boldsymbol{q}_n)$. In a first step probabilistic inversion is accomplished by

estimating $\boldsymbol{\theta}_{\boldsymbol{X}}$ with the data $\langle \boldsymbol{y}_{\neq i_0} \rangle$. Similarly to Eq. (4.35), the resulting posterior $\pi(\boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_{\neq i_0} \rangle)$ can be translated into a mixture distribution

$$\pi\big(\boldsymbol{x}_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle\big) = \int_{\mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{X}}}} f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_{i_0} | \boldsymbol{\theta}_{\boldsymbol{X}})\, \pi\big(\boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_{\neq i_0} \rangle\big)\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}. \tag{4.36}$$

It represents the uncertainty in $\boldsymbol{x}_{i_0}$ following the analysis of $\langle \boldsymbol{y}_{\neq i_0} \rangle$ but prior to analyzing $\boldsymbol{y}_{i_0}$. Thereupon the second stage of the filtering program consists in utilizing Eq. (4.36) as a Bayesian prior for inferring $\boldsymbol{x}_{i_0}$ by inverting $\boldsymbol{y}_{i_0}$. Bayesian updating yields the posterior distribution $\pi(\boldsymbol{x}_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle, \boldsymbol{y}_{i_0}) \propto f_{\boldsymbol{E}}(\boldsymbol{y}_{i_0} - \mathcal{M}(\boldsymbol{x}_{i_0}, \boldsymbol{d}_{i_0}); \boldsymbol{\Sigma}_{i_0})\, \pi(\boldsymbol{x}_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle)$.

Information-wise, the estimation of $\boldsymbol{\theta}_{\boldsymbol{X}}$ has been initially based on the data $\langle \boldsymbol{y}_{\neq i_0} \rangle$, its conditional distributions $f_{\boldsymbol{E}}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i)$ for $i \neq i_0$, the structural knowledge $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})$ and the parametric prior $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$. While inheriting the obtained information about $\boldsymbol{\theta}_{\boldsymbol{X}}$ by means of Eq. (4.36), the observation $\boldsymbol{y}_{i_0}$ has been eventually inverted for $\boldsymbol{x}_{i_0}$.

4.6.3 Multilevel inversion

A full hierarchical analysis constitutes the third type of estimation. By formulating the joint posterior Eq. (4.32) of the collectivity of unknowns $(\langle \boldsymbol{x}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}})$ and marginalizing over nuisance $(\langle \boldsymbol{x}_{\neq i_0} \rangle, \boldsymbol{\theta}_{\boldsymbol{X}})$, the posterior distribution of the QoI $\boldsymbol{x}_{i_0}$ can be written as

$$\pi\big(\boldsymbol{x}_{i_0} | \langle \boldsymbol{y}_i \rangle\big) = \int_{\mathcal{D}_{\boldsymbol{x}}^{n-1}} \int_{\mathcal{D}_{\boldsymbol{\theta}_{\boldsymbol{X}}}} \pi\big(\langle \boldsymbol{x}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle\big)\, \mathrm{d}\langle \boldsymbol{x}_{\neq i_0} \rangle\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}, \tag{4.37}$$

where $\mathrm{d}\langle \boldsymbol{x}_{\neq i_0} \rangle = \mathrm{d}\boldsymbol{x}_1 \dots \mathrm{d}\boldsymbol{x}_{i_0-1}\, \mathrm{d}\boldsymbol{x}_{i_0+1} \dots \mathrm{d}\boldsymbol{x}_n$. Note that when the joint posterior Eq. (4.32) is computed, other marginals than Eq. (4.37) can be extracted similarly.

In terms of estimating $\boldsymbol{x}_{i_0}$, the structure of the posterior Eq. (4.37) reveals that all the different pieces of information have been "optimally" combined during a joint learning process. From an informational point of view, the total data $\langle \boldsymbol{y}_i \rangle$, their conditional distributions $f_{\boldsymbol{E}}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i)$, the structural knowledge $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})$ and the hyperprior $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ have been completely synthesized. This implies that inferring $\boldsymbol{x}_{i_0}$ "borrows" information encoded in the observations $\langle \boldsymbol{y}_{\neq i_0} \rangle$. A DAG-based visualization of the underlying flow of information is provided in Fig. 4.3. The deeper reason for borrowing strength to happen is the partial reducibility of the uncertainty model Eq. (4.8), i.e. the exchangeability of aleatory variables $\langle \boldsymbol{x}_i \rangle$.

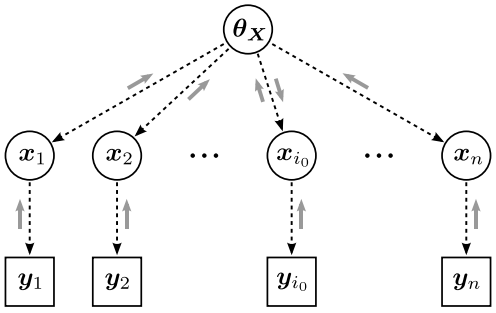


Figure 4.3: Optimal combination of information. A Bayesian network representation of probabilistic inversion is shown. Known (□) and unknown (○) quantities are related by probabilistic (⇢) relations. The "upstream" (⟵) and "downstream" (⟶) flow of information towards a specific $\boldsymbol{x}_{i_0}$ is indicated. This is a form of borrowing strength.

4.7 Bayesian computations

Generally Bayesian posteriors feature an analytic closed-form expression only on a rare occasion. Specifically this applies to posteriors of the form Eqs. (4.13), (4.21) and (4.28). Notwithstanding the above, posteriors can be explored by means of Markov chain Monte Carlo (MCMC) [79, 80]. Principally this readily refers to posteriors stemming from multilevel inversion. The Metropolis-Hastings (MH) algorithm and the Gibbs sampler are prototypical MCMC techniques. In Section 4.7.1 we will review the MH algorithm and discuss classical MCMC key issues in Section 4.7.2. Additional computational key challenges posed by Bayesian multilevel model calibration will be discussed in Section 4.7.3. Some more sophisticated MCMC samplers that are suitable in a multilevel-context are surveyed in Section 4.7.4.

4.7.1 The Metropolis-Hastings algorithm

MCMC is based on constructing an ergodic Markov chain such that its invariant distribution equals the posterior. Let $\pi(\boldsymbol{q})$ be the prior and $\pi(\boldsymbol{q}|\langle \boldsymbol{y}_i \rangle)$ the posterior density of some QoI $\boldsymbol{q}$. A Markov chain with equilibrium distribution $\pi(\boldsymbol{q}|\langle \boldsymbol{y}_i \rangle)$ is generated by initializing at $\boldsymbol{q}^{(0)}$ and repetitively proceeding as follows. Given a state $\boldsymbol{q}^{(t)}$ that the Markov chain has taken on in some iteration, in the following iteration a candidate state $\boldsymbol{q}^{(\star)} \sim P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})$ is randomly sampled from a proposal distribution $P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})$. In the MH correction step the proposed state is approved as the new state $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(\star)}$ of the Markov chain with probability

$$\alpha\left(\boldsymbol{q}^{(\star)}, \boldsymbol{q}^{(t)}\right) = \min\left(1, \frac{\pi(\boldsymbol{q}^{(\star)}|\langle \boldsymbol{y}_i \rangle)\, P(\boldsymbol{q}^{(t)}|\boldsymbol{q}^{(\star)})}{\pi(\boldsymbol{q}^{(t)}|\langle \boldsymbol{y}_i \rangle)\, P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})}\right). \tag{4.38}$$

Otherwise the proposal will be rejected, i.e. the Markov chain remains in its state $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}$ of the preceding iteration. It is important to note that due to the MH acceptance probability Eq. (4.38), the algorithm calls for the computation of posterior ratios only. Thus for MCMC sampling the scale factors in Eqs. (4.13), (4.21) and (4.28) can be dropped and only unscaled posterior densities have to be evaluated.

Random walk Metropolis sampling rests upon local proposals, e.g. candidate states are sampled from a Gaussian distribution $\boldsymbol{q}^{(\star)} \sim \mathcal{N}(\boldsymbol{q}^{(\star)}; \boldsymbol{q}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{q}})$ that is centered around the current state $\boldsymbol{q}^{(t)}$. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{q}}$ determines the "stepsizes" of the algorithm. Independence MH sampling is based on nonlocal proposals whose distribution $\boldsymbol{q}^{(\star)} \sim P(\boldsymbol{q}^{(\star)})$ is independent of $\boldsymbol{q}^{(t)}$, e.g. sampling candidate states from the prior $\boldsymbol{q}^{(\star)} \sim \pi(\boldsymbol{q}^{(\star)})$ or from some suitable approximation of the posterior $\boldsymbol{q}^{(\star)} \sim \hat{\pi}(\boldsymbol{q}^{(\star)}|\langle \boldsymbol{y}_i \rangle)$.

4.7.2 Classical key challenges

The performance of MCMC methods is governed by the mixing properties of the underlying Markov chain, i.e. the speed of convergence of the Markov chain towards the targeted posterior. As to which degree MCMC samples are autocorrelated has a determining influence on the convergence speed and their quality as posterior representatives. Hence MCMC algorithms are designed and tuned in pursuit of rapid mixing. Depending on the specific problem at hand, this may be a tricky business which requires to employ and combine sophisticated and highly specialized sampling schemes. Typically MCMC sampling calls for a high number of program iterations which in turn demands a high number of forward model runs for evaluating the likelihood function in the MH correction Eq. (4.38). Beyond that, careful convergence diagnostics are of particular importance for MCMC methods. One has to assess when the Markov chain has reached its stationary distribution, i.e. when it has lost any dependence on its initialization. Even though there are advanced convergence test [81, 82], e.g. Gelman-Rubin diagnostics for multiple over-dispersed chains [83, 84], we remark that from a pessimistic point of view any convergence diagnostic is heuristics [85]. Furthermore MCMC suffers from difficulties in exploring high-dimensional and multimodal posteriors.

4.7.3 Multilevel-related challenges

Multilevel posteriors can be readily sampled by means of classical MCMC techniques as they are commonly applied in "simple" Bayesian inversion. However, on top of the classical bottlenecks that were discussed above, one is faced with multilevel-specific MCMC challenges. The posteriors Eqs. (4.13) and (4.21), which are appertain to the joint and the marginal variant of multilevel calibration, are different in nature. Accordingly, sampling these posteriors pose different computational burdens. The former requires a sampling scheme that performs efficiently in high-dimensional parameter spaces, whereas the latter suffers from computing the integrated likelihood Eq. (4.20). Similarly the posterior Eq. (4.28) of the "perfect" data model imposes forward uncertainty quantification for the computation of the likelihood Eq. (4.27).

Likelihood functions of the form Eqs. (4.23) and (4.30) suffer from another severe difficulty. It is well-known that statistical estimations of the likelihood ratio introduce an additional random component into the Markov chain transition kernel [86, 87]. Consequently the steady-state distribution of the chain may be modified. Therefore free parameters of the algorithm have to be chosen endeavoring high *posterior fidelity*, i.e. the degree as to which the induced long-run distribution conforms with the true posterior [59, 60].

4.7.4 Advanced MCMC samplers

Summarized Bayesian multilevel model calibration requires an enormous number of forward model runs. Therefore in the statistical literature a wide range of advanced MCMC techniques, dedicated to posterior exploration in classical hierarchical models, have been devised. Some enhanced Gibbs sampling methods in this context are reviewed in [79] and references therein. However, in view of engineering problems they may not

meet the challenges those applications usually pose. This is due to the inescapable "blackbox" character of the forward solver and nonconjugacy. Generally not all of the parameters will have full conditionals of a standard form that can be easily sampled. Despite that this paper does not focus on computational facets of uncertainty quantification, a short outlook on potentially efficient MCMC implementations is given.

Data augmentation is a powerful MCMC technique that aims at enhancing the numerical efficiency of posterior computation by introducing missing data as auxiliary variables [88, 89]. Note that the joint posterior Eq. (4.13) can be seen as an augmented form of the marginal one in Eq. (4.21). Thus data augmentation naturally emerges in the context of Bayesian multilevel inversion. It has been beneficially applied for solving multilevel inverse problems within the domain of aerospace engineering [59, 60]. Vice versa, there are dedicated MCMC schemes for directly computing marginalized posteriors of the form Eq. (4.21), e.g. MC within Metropolis sampling [55, 86] or pseudo-marginalization [90]. The Hamiltonian Monte Carlo (HMC) algorithm is a sampler whose performance is remarkably efficient in high-dimensional parameter spaces and for highly correlated posteriors [91, 92]. Since multilevel models are higher-dimensional and correlated by definition, the HMC is a promising MCMC candidate in this context. Yet the HMC still occurs to be highly underacknowledged in Bayesian inference in general and for hierarchical models in particular.

4.8 Numerical case studies

In order to illustrate the power and versatility of the devised framework we conduct a selection of computer experiments. This shall be seen as a proof of concept and benchmark of the proposed methodology in the context of engineering applications. A system of identically designed structural components functions as the basis for probing a range of experimental scenarios. Specifically we deal with an ensemble of simply supported beams that are tested in a series of three-point bending experiments. By multilevel analysis of measured beam deflections we highlight how different inferential goals, e.g. probabilistic inversion, residual calibration or optimal combination of information, can be achieved in the presence of material variability and uncertainties in the experimental setup. Keeping deterministic modeling simple and intuitive will allow us to focus on uncertainty quantification aspects that are the essential subject matter of this research. Incidentally we learn about the computational obstacles that must be overcome when aiming at "real-world" applications.

The forward problem will be shortly introduced in Section 4.8.1. Around this submodel, that covers the deterministic features of the system, Bayesian multilevel models will be built to capture uncertainty and variability. Probabilistic inversion, i.e. deducing the material variability throughout an ensemble of similar specimens, will be tackled in Section 4.8.2. The subsequent Section 4.8.3 will deal with residual model calibration. In Section 4.8.4 the impact of prescribed uncertainties in the test conditions will be investigated. In Section 4.8.5 borrowing strength will be utilized in order to ideally estimate the material characteristics of a single specimen by using information obtained from the other specimens.

4.8.1 Mechanical model

The system under consideration is an ensemble of identically manufactured beams $i = 1, \dots, n$ with well-known lengths L_i and rectangular cross sections with widths b_i and heights h_i. Yet the completed beams are only similar in the sense that we assume variability in the elastic moduli E_i across the ensemble, e.g. due to slight irregularities in the fabrication process. For each single beam i the Young's modulus E_i is assumed to be constant along the main beam axis. The deflections $\tilde{v}_i(s_{i,j})$ of a simply supported beam i under a concentrated point load F_i at midspan can be easily derived in Euler-Bernoulli beam theory. For positions $s_{i,j}$ along the beam axis with $0 \leq s_{i,j} \leq L_i/2$ and $j = 1, \dots, n_i$ the deflections follow as

$$\tilde{v}_i(s_{i,j}) = \frac{F_i s_{i,j}}{48 E_i I_i} \left(3L_i^2 - 4s_{i,j}^2\right), \quad \text{for } 0 \leq s_{i,j} \leq L_i/2, \tag{4.39}$$

where the moment of inertia is given as $I_i = b_i h_i^3/12$. Likewise a symmetric expression holds for positions $s_{i,j}$ along the main axis with $L_i/2 \leq s_{i,j} \leq L_i$. A single simply supported beam is visualized in Fig. 4.4.

Together with its symmetric counterpart, the algebraic formula Eq. (4.39) constitutes the deterministic submodel of the system under consideration. When a load F_i is applied to a beam i with physical dimensions $\boldsymbol{l}_i = (L_i, b_i, h_i)$ and an elastic modulus E_i, these relations predict the deflections $\tilde{\boldsymbol{v}}_i = (\tilde{v}_i(s_{i,1}), \dots, \tilde{v}_i(s_{i,n_i}))$ at positions $\boldsymbol{s}_i = (s_{i,1}, \dots, s_{i,n_i})$. We denote this as

$$\tilde{\boldsymbol{v}}_i = \mathcal{M}(E_i, F_i, \boldsymbol{l}_i, \boldsymbol{s}_i). \tag{4.40}$$

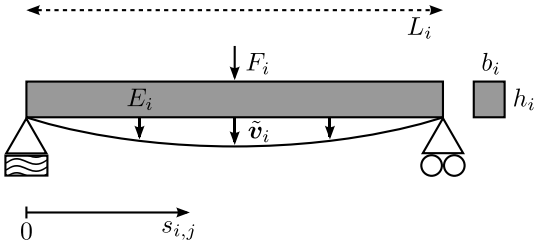


Figure 4.4: A simply supported beam.

When beam deflections are measured in three-point bending tests for each member $i = 1, \dots, n$ in the population, multilevel inversion allows for optimal data analysis in experimental situations where the inputs of Eq. (4.40) are subject to uncertainty.

4.8.2 Probabilistic inversion

We begin with Bayesian probabilistic inversion, on the basis of which we demonstrate how one can quantify the material variability within the ensemble of beams in a series of bending tests. A numerical experiment is therefore set up as follows. We consider a number of $n = 100$ beams with well-known dimensions $L_i = 1\,\mathrm{m}$ and $b_i = h_i = 10\,\mathrm{cm}$. Beams are subjected to concentrated loads $F_i = 30\,\mathrm{kN}$ that are applied at midspan. For $i = 1, \dots, 100$ Young's moduli E_i are independently sampled from a lognormal distribution $\mathcal{LN}(E_i|\mu_E, \sigma_E)$ with mean $\mu_E = 15\,\mathrm{GPa}$ and standard deviation $\sigma_E = 3\,\mathrm{GPa}$. This corresponds to a coefficient of variation $c_E = 20\,\%$. After having set up the experiment, the hyperparameters $\boldsymbol{\theta}_E = (\mu_E, \sigma_E)$ as well as beam-specific moduli E_i will be treated as unknowns. At $n_i = 3$ positions $\boldsymbol{s}_i = (s_{i,1}, s_{i,2}, s_{i,3})$ with $s_{i,1} = 25\,\mathrm{cm}$, $s_{i,2} = 50\,\mathrm{cm}$ and $s_{i,3} = 75\,\mathrm{cm}$ beam deflections $\tilde{\boldsymbol{v}}_i = (\tilde{v}_i(s_{i,1}), \tilde{v}_i(s_{i,2}), \tilde{v}_i(s_{i,3}))$ are computed according to Eq. (4.39). In order to take measurement uncertainty and forward model imperfection into account, we perturb the predictions $\tilde{\boldsymbol{v}}_i$ with noise terms $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \varepsilon_{i,2}, \varepsilon_{i,3})$. Those terms are independently sampled from Gaussian distributions $\mathcal{N}(\boldsymbol{\varepsilon}_i; \mathbf{0}, \boldsymbol{\Sigma}_i)$ with $\boldsymbol{\Sigma}_i = \sigma_i^2 \boldsymbol{I}_3$ and $\sigma_i = 0.1\,\mathrm{mm}$. Eventually $\boldsymbol{v}_i = \tilde{\boldsymbol{v}}_i + \boldsymbol{\varepsilon}_i$ represent the pseudo data that will become analyzed with respect to the QoI $\boldsymbol{\theta}_E = (\mu_E, \sigma_E)$.

In many circumstances expert knowledge about the QoI $\boldsymbol{\theta}_E$ is available prior to analyzing the data. This knowledge can be accounted for by eliciting a suitable prior distribution $\pi(\boldsymbol{\theta}_E)$. Herein we employ a proper Bayesian prior $\pi(\boldsymbol{\theta}_E) = \pi(\mu_E)\,\pi(\sigma_E)$ with independent marginals. As measured in units of GPa those marginals are given as uniform distributions $\pi(\mu_E) = \mathcal{U}(0, 100)$ and $\pi(\sigma_E) = \mathcal{U}(0, 30)$. This is supposed to represent an experimental situation where one cannot elicit informative priors, nonetheless one is confident enough to assign this weakly informative and flat prior with its upper and lower bounds.

Ultimately probabilistic inversion can be summarized as the estimation of the QoI $\boldsymbol{\theta}_{\boldsymbol{X}} \equiv \boldsymbol{\theta}_E$ with the deflection measurements $\langle \boldsymbol{y}_i \rangle \equiv \langle \boldsymbol{v}_i \rangle$. Beam-specific Young's moduli $\langle \boldsymbol{x}_i \rangle \equiv \langle E_i \rangle$, that are not of immediate inferential interest, are considered nuisance to that end. Experimental conditions $\langle \boldsymbol{d}_i \rangle \equiv \langle (F_i, \boldsymbol{l}_i, \boldsymbol{s}_i) \rangle$, that the experiments where subject to, and prediction error models $\langle \boldsymbol{\Sigma}_i \rangle$ are assumed to be known. The distributions $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}}) \equiv \mathcal{LN}(E_i|\mu_E, \sigma_E)$ and $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}) \equiv \pi(\boldsymbol{\theta}_E)$ represent the available structural and parametric prior knowledge, respectively. The emerging posterior will be of the form $\pi(\boldsymbol{\theta}_{\boldsymbol{X}}|\langle \boldsymbol{y}_i \rangle) \equiv \pi(\boldsymbol{\theta}_E|\langle \boldsymbol{v}_i \rangle)$. It can be directly sampled or accessed via the QoI-marginals of the joint posterior $\pi(\langle \boldsymbol{x}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}}|\langle \boldsymbol{y}_i \rangle) \equiv \pi(\langle E_i \rangle, \boldsymbol{\theta}_E|\langle \boldsymbol{v}_i \rangle)$. A DAG corresponding to probabilistic inversion is provided in Fig. 4.2(b).

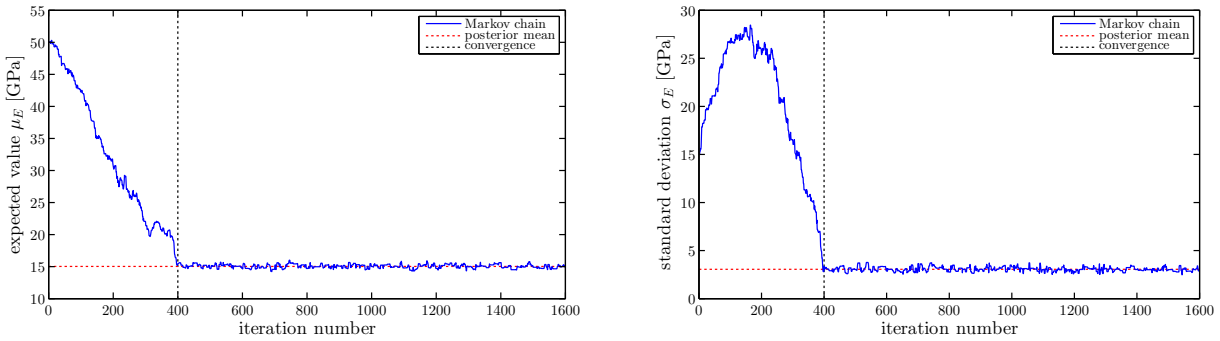
4.8.2.1 MCMC

Generally we employ a joint rather than a marginal problem formulation. For the fidelity reasons that were discussed in Section 4.7.3 this allows for exact posterior computation where an approximation is only introduced in as much as MCMC sampling is concerned. Moreover a joint posterior features a richer structure which will provide new insights into multilevel inversion. All computations will be serially done on a contemporary Intel Xeon CPU.

The joint posterior $\pi(\langle E_i \rangle, \boldsymbol{\theta}_E|\langle \boldsymbol{v}_i \rangle)$ is sampled by means of a blockwise random walk Metropolis algorithm. A practical problem of random walk samplers in high dimension is to carefully tune the proposal distribution. For complex multivariate posterior distributions this is a cumbersome procedure that poses severe difficulties. However, in multilevel inversion one can advantageously exploit the "symmetry" of the problem in the latent variables. Assuming that separate inverse problems i with $1 \leq i \leq n$ are not severely ill-posed, latent variables of the same uncertainty type are expected to behave similarly in the sense that their marginal posteriors resemble one another. Moreover, due to the indirectness of borrowing strength, their mutual correlations are expected

to be rather small. Along these lines the "effective dimensionality" is lower than the number of unknowns suggests. This discussion motivates that MCMC updates are done in blocks $\langle E_i \rangle$ and (μ_E, σ_E). We find that with Gaussian jumping distributions the algorithm can be easily tuned in such a way that blockwise acceptance rates range between 20 % and 40 %. Avoiding lengthy convergence times in high-dimensional problems requires smart initialization, too. Again we proceed by exploiting the structure of the multilevel system. The block $\langle E_i \rangle$ is initialized with solutions of separate inverse problems, while two-stage estimates are used in the hyperparameter block (μ_E, σ_E).

In order to assure duly completed posterior exploration we perform a number of convergence checks. The algorithm is initialized in regions of the parameter space that had not been visited before and the convergence behavior of the Markov chain is monitored. We detect that the chain eventually reaches the same posterior modes again. In Fig. 4.5 trace plots of a converging Markov chain are shown for its μ_E and σ_E components. They have been initialized at $\mu_E^{(0)} = 50\,\text{GPa}$ and $\sigma_E^{(0)} = 15\,\text{GPa}$, i.e. in the middle of their priorly admissible intervals. While the mean hyperparameter μ_E directly converges as shown in Fig. 4.5(a), we observe a different behavior for the spread hyperparameter σ_E. From Fig. 4.5(b) it can be seen that the latter chain tends to higher values prior to attraction towards the posterior mean. For the given initialization this is a systematic effect that indicates a posterior correlation in the hyperparameters (μ_E, σ_E). Eventually the Markov chain converges within ca. 400 MCMC iterations. Apart from such visual inspections we generally rely on Gelman-Rubin diagnostics for parallel chains [83, 84].

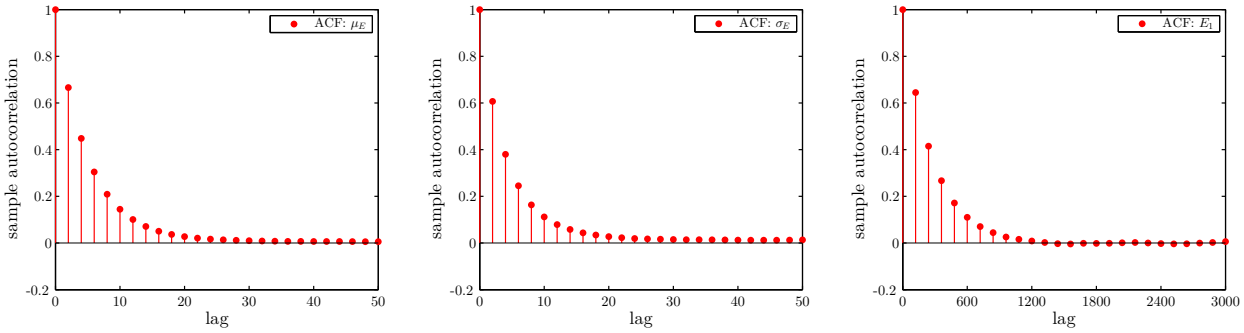


(a) Convergence of μ_E.

(b) Convergence of σ_E.

Figure 4.5: Trace plots of a converging Markov chain. For $n = 100$ the converging Markov chain is shown for μ_E in (a) and for σ_E in (b). Being initialized at $\mu_E^{(0)} = 50\,\text{GPa}$ and $\sigma_E^{(0)} = 15\,\text{GPa}$ the Markov chain converges within ca. 400 MCMC iterations. In equilibrium the Markov chain samples the posterior around its mean.

In Fig. 4.6 the MCMC sample autocorrelations are plotted for the QoI (μ_E, σ_E) and for an intermediate variable E_i with $i = 1$. It can be seen how the autocorrelation function (ACF) drops until it becomes indistinguishable from zero. This behavior governs the quality of the sample as a posterior representative. Especially the ACF of E_i shown in Fig. 4.6(c) motivates more efficient updating schemes in future research.



(a) Autocorrelation of μ_E.

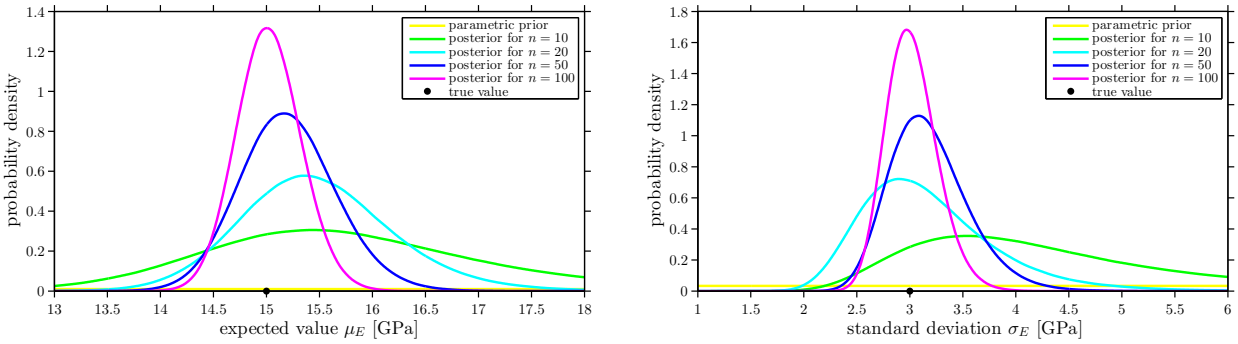
(b) Autocorrelation of σ_E.

(c) Autocorrelation of E_1.

Figure 4.6: Sample autocorrelation functions. For a run with $n = 100$ the MCMC sample autocorrelation function is plotted for μ_E in (a), for σ_E in (b) and for E_1 in (c). The sample autocorrelation determines the effective MCMC sample size.

4.8.2.2 Results: Posterior marginals

We analyze the data $\langle \boldsymbol{v}_i \rangle_{1 \le i \le 100}$ as well as its subconfigurations $\langle \boldsymbol{v}_i \rangle_{1 \le i \le 10}$, $\langle \boldsymbol{v}_i \rangle_{1 \le i \le 20}$ and $\langle \boldsymbol{v}_i \rangle_{1 \le i \le 50}$. This allows to assess how the number of experiments n influences the identification of the QoI. For each of the runs $N = 10^7$ MCMC iterations are performed. As a general rule we discard the initial 1% of the total number of iterations of each Markov chain as a burn-in period. The total algorithm runtime adds up to $t = 3.85\,\mathrm{h}$ for $n = 10$ and to $t = 4.66\,\mathrm{h}$ for $n = 100$. The resulting posterior marginals of μ_E and σ_E are shown in Fig. 4.7. A statistical summary of these marginals can be found in Table 4.1, where the mean, mode, standard deviation (SD) and coefficient of variation (CV) are listed. With increasing number of processed experiments n, Bayesian point estimates (mean, mode) approach the true values $\mu_E = 15\,\mathrm{GPa}$ and $\sigma_E = 3\,\mathrm{GPa}$ while measures of estimation uncertainty (SD, CV) expectedly decrease.



(a) Posterior marginal of μ_E.

(b) Posterior marginal of σ_E.

Figure 4.7: Posterior marginals of the QoI. Corresponding to various numbers of experiments n, the marginal posterior densities of μ_E and σ_E are shown in (a) and (b), respectively. For increasing n, the posterior uncertainty in estimating the QoI $\boldsymbol{\theta}_E = (\mu_E, \sigma_E)$ with $\mu_E = 15\,\mathrm{GPa}$ and $\sigma_E = 3\,\mathrm{GPa}$ steadily decreases.

Table 4.1: Summary of the QoI posterior marginals.

	μ_E [GPa]			[ ]	σ_E [GPa]			[ ]
	Mean	Mode	SD	CV	Mean	Mode	SD	CV
$n = 10$	15.98	15.43	2.06	0.13	4.73	3.54	3.55	0.75
$n = 20$	15.48	15.36	0.74	0.05	3.18	2.90	0.65	0.20
$n = 50$	15.20	15.17	0.46	0.03	3.17	3.08	0.37	0.12
$n = 100$	15.02	15.00	0.30	0.02	3.02	2.97	0.24	0.08

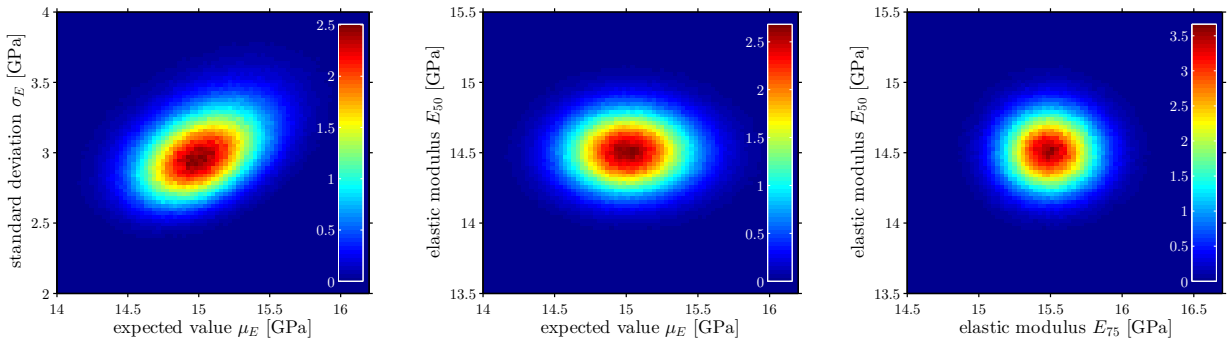
4.8.2.3 Results: Two-dimensional posteriors

Showing posterior marginals may hide possibly existing dependency structures or the lack thereof. Those constitute a substantial result of Bayesian data analysis, though. Hence Fig. 4.8 shows two-dimensional posteriors where interesting correlation properties were discovered. The two-dimensional posterior of (μ_E, σ_E) is plotted in Fig. 4.8(a). According to the posterior probability model these two parameters are correlated with a linear Pearson coefficient of correlation $r_{\mu_E, \sigma_E} = 0.40$. Note that these parameters were assumed to be independent in accord with their prior model. The joint posterior Eq. (4.32) can also feature a correlation between hyperparameters and experiment-specific parameters. In Figs. 4.8(b) and 4.8(c) the two-dimensional posteriors of (μ_E, E_i) and (E_j, E_i) with $i = 50$ and $j = 75$ are imaged.

4.8.3 Residual calibration

There are situations where the strong assumption of known residual variances $\boldsymbol{\Sigma}_i = \sigma_i^2 \boldsymbol{I}_3$ is somewhat restrictive. Thus we generalize multilevel inversion as in Section 4.8.2 by treating $\sigma_{\mathcal{E}} \equiv \sigma_i$ as a global unknown. In units of mm the corresponding parametric prior is set to a uniform distribution $\pi(\sigma_{\mathcal{E}}) = \mathcal{U}(0, 0.5)$. Otherwise the experimental setup of probabilistic inversion is used.

The standard deviation $\sigma_{\mathcal{E}}$ of the residual model $\mathcal{N}(\boldsymbol{\varepsilon}_i | \boldsymbol{0}, \sigma_{\mathcal{E}}^2 \boldsymbol{I}_3)$ is introduced as an extra unknown in the model Eq. (4.31) and in the posterior Eq. (4.32). Consequently the joint prior is given as $\pi(\langle E_i \rangle, \mu_E, \sigma_E, \sigma_{\mathcal{E}}) =$



(a) 2D posterior of (μ_E, σ_E). (b) 2D posterior of (μ_E, E_{50}). (c) 2D posterior of (E_{75}, E_{50}).

Figure 4.8: 2D posteriors of (μ_E, σ_E), (μ_E, E_{50}) and (E_{75}, E_{50}). The two-dimensional posteriors of (μ_E, σ_E), (μ_E, E_{50}) and (E_{75}, E_{50}) are shown. Being priorly independent the components μ_E and σ_E are seen to be correlated a posteriori. The linear Pearson coefficient of correlation amounts to $r_{\mu_E,\sigma_E} = 0.40$.

$\pi(\sigma_\mathcal{E})\,\pi(\mu_E)\,\pi(\sigma_E)\prod_{i=1}^{n}\mathcal{LN}(E_i|\mu_E,\sigma_E)$. For the joint likelihood function one has $\mathcal{L}(\langle E_i\rangle,\sigma_\mathcal{E};\langle \boldsymbol{v}_i\rangle) = \prod_{i=1}^{n}\mathcal{N}(\boldsymbol{v}_i|\mathcal{M}(E_i,F_i,\boldsymbol{l}_i,\boldsymbol{s}_i),\sigma_\mathcal{E}^2\boldsymbol{I}_3)$. Brought together this leads to a joint posterior density that has the shape $\pi(\langle E_i\rangle,\mu_E,\sigma_E,\sigma_\mathcal{E}|\langle \boldsymbol{v}_i\rangle) \propto \mathcal{L}(\langle E_i\rangle,\sigma_\mathcal{E};\langle \boldsymbol{v}_i\rangle)\,\pi(\langle E_i\rangle,\mu_E,\sigma_E,\sigma_\mathcal{E})$.

We sample from this posterior by appending a block for the additional unknown $\sigma_\mathcal{E}$ in the MCMC updating scheme. In order to assess the influence of the amount of data on the final results, independent runs are performed for $n = 10$, 20, 50 and 100. In Fig. 4.9 the relevant posterior marginals for the inference of the residual model $\sigma_\mathcal{E}$ are shown. A short summary of the these marginals is provided in Table 4.2. The higher the number of analyzed experiments n, the better the true value $\sigma_\mathcal{E} = 0.1\,\text{mm}$ has been revealed. This proves that one can indeed estimate the parameters of the prediction error model in the context of multilevel calibration. If this is not of interest for its own sake, it still avoids the requirement of perfect knowledge of the error variance. In addition we observed that introducing an uncertainty in the residual model hardly affects the inference of the QoI in probabilistic inversion.

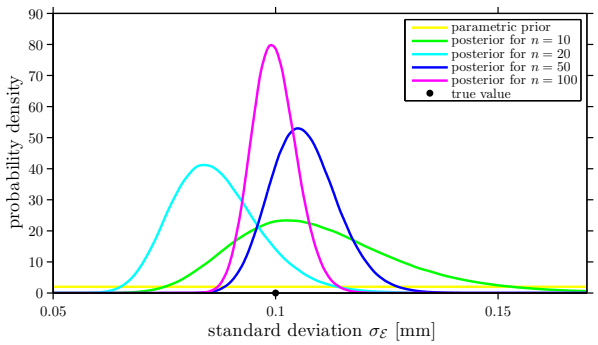


Figure 4.9: Posterior marginals of $\sigma_\mathcal{E}$. The marginal posterior of $\sigma_\mathcal{E}$ is shown for different numbers of data n.

Table 4.2: Summary of the $\sigma_\mathcal{E}$-marginals.

	$\sigma_\mathcal{E}$ [10^{-5} m]			[ ]
	Mean	Mode	SD	CV
$n = 10$	11.00	10.23	1.90	0.17
$n = 20$	8.68	8.38	1.01	0.12
$n = 50$	10.65	10.50	0.77	0.07
$n = 100$	9.97	9.90	0.50	0.05

4.8.4 Uncertain conditions

In the following we describe an experimental situation where the inference of the QoI $\boldsymbol{\theta}_E$ is hampered by additional uncertainties in the experimental conditions. Experimental conditions are formally treated as nuisance parameters with prescribed uncertainties. More specifically, we do not assume that the loads F_i are perfectly known anymore. In contrast, we assume that they are $\boldsymbol{\zeta}_i$-type variables, i.e. they are uncertain yet they follow a known distribution. This represents a well-known situation where the loads F_i that the testing machine actually applies can only be imprecisely adjusted. In fact, while a targeted load in each experiment is chosen, the physically realized load F_i may be uncertain. This is accounted for by a prescribed distribution $\mathcal{N}(F_i;\mu_{F_i},\sigma_{F_i}^2)$ where μ_{F_i} is the targeted load and σ_{F_i} represents the degree of uncertainty that is inherent to the test machinery.

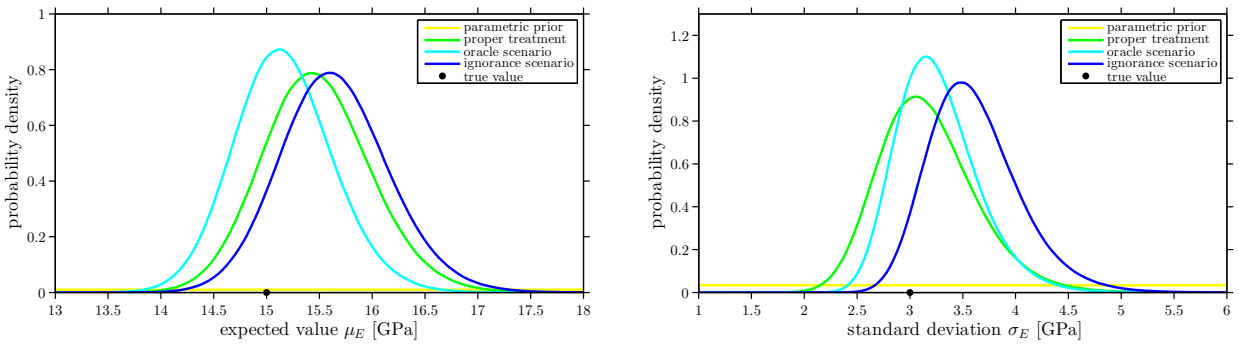
The setup for conducting a numerical experiment is similar to the one specified in Section 4.8.2. For $n = 50$ beams we set the beam dimensions $\boldsymbol{l}_i$ and measurement positions $\boldsymbol{s}_i$ as before. Elastic moduli E_i are randomly drawn from $\mathcal{LN}(E_i|\mu_E,\sigma_E)$ as previously detailed. In contrast to plain probabilistic inversion, for

$i = 1, \dots, n$ experiment-specific loads F_i are independently sampled from normal distributions $\mathcal{N}(F_i; \mu_{F_i}, \sigma^2_{F_i})$ with $\mu_{F_i} = 30\,\text{kN}$ and $\sigma_{F_i} = 3\,\text{kN}$. This equates to a coefficient of variation $c_{F_i} = 10\,\%$. Note that such a high degree of uncertainty is unlikely to be encountered in a real-case experiment. It is used here to accentuate the results presented below, though. The realized loads F_i will be treated as unknowns whereas the hyperparameters $\boldsymbol{\theta}_{F_i} = (\mu_{F_i}, \sigma_{F_i})$, i.e. the targeted load and its uncertainty, will be treated as knowns. In accordance with Eq. (4.39) synthetic measurements $\boldsymbol{v}_i = \tilde{\boldsymbol{v}}_i + \boldsymbol{\varepsilon}_i$ are generated again. The prior distribution $\pi(\boldsymbol{\theta}_E) = \pi(\mu_E)\,\pi(\sigma_E)$ is also chosen as previously stated.

The problem of probabilistic inversion under additional prescribed nuisance reads as follows. The hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}} \equiv \boldsymbol{\theta}_E$ are the QoI whereas experiment-specific unknowns $\langle \boldsymbol{x}_i \rangle \equiv \langle E_i \rangle$ and $\langle \boldsymbol{\zeta}_i \rangle \equiv \langle F_i \rangle$ are considered nuisance. With measurements $\langle \boldsymbol{y}_i \rangle \equiv \langle \boldsymbol{v}_i \rangle$ the QoI can be inferred. Experimental-specific knowns consist of the hyperparameters $\langle \boldsymbol{\theta}_{\boldsymbol{Z}_i} \rangle \equiv \langle \boldsymbol{\theta}_{F_i} \rangle$, the experimental conditions $\langle \boldsymbol{d}_i \rangle \equiv \langle (\boldsymbol{l}_i, \boldsymbol{s}_i) \rangle$ and the residual covariances $\langle \boldsymbol{\Sigma}_i \rangle$. Parametric Bayesian prior knowledge is given by $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}) \equiv \pi(\boldsymbol{\theta}_E)$ whereas $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}}) \equiv \mathcal{LN}(E_i|\mu_E, \sigma_E)$ and $f_{\boldsymbol{Z}}(\boldsymbol{\zeta}_i; \boldsymbol{\theta}_{\boldsymbol{Z}_i}) \equiv \mathcal{N}(F_i; \mu_{F_i}, \sigma^2_{F_i})$ are structural prior distributions. Within a joint approach a posterior of the form $\pi(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle) \equiv \pi(\langle E_i \rangle, \langle F_i \rangle, \boldsymbol{\theta}_E | \langle \boldsymbol{v}_i \rangle)$ arises. Eventually one is interested in the QoI-marginals $\pi(\boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle) \equiv \pi(\boldsymbol{\theta}_E | \langle \boldsymbol{v}_i \rangle)$ only. A DAG corresponding to this experimental situation is shown in Fig. 4.2(c).

4.8.4.1 Results: Hyperparameters

We sample the joint posterior $\pi(\langle E_i \rangle, \langle F_i \rangle, \boldsymbol{\theta}_E | \langle \boldsymbol{v}_i \rangle)$ where nuisance variables $\langle F_i \rangle$ are explicitly accounted for. In a blockwise manner MCMC sweeps are accomplished for (μ_E, σ_E), $\langle E_i \rangle$ and $\langle F_i \rangle$ which constitute different blocks. Blockwise proposal distributions are again adjusted in order to obtain acceptance rates in between 20 % and 40 %. Each F_i in the block $\langle F_i \rangle$ is initialized at $F_i^{(0)} = \mu_{F_i}$, i.e. the structural prior mean. Other than that initialization, convergence checks and burn-in are accomplished as before. For $N = 10^7$ MCMC iterations the total computation time amounts to $t = 7.18\,\text{h}$. The resulting posterior marginals of μ_E and σ_E can be seen in Fig. 4.10. A statistical summary is provided in Table 4.3 where the mean, mode, SD and CV of the marginals are itemized.



(a) Posterior marginal of μ_E.

(b) Posterior marginal of σ_E.

Figure 4.10: Posterior marginals of the QoI. The marginal posteriors of μ_E and σ_E are provided in (a) and (b), respectively. Three experimental scenarios are investigated: the proper treatment of the additional uncertainty, an idealized situation where one would precisely know the loads, and the case of a parsimonious model where the uncertainty remains unrecognized.

Table 4.3: Summary of the QoI posterior marginals.

	μ_E [GPa]			[ ]	σ_E [GPa]			[ ]
	Mean	Mode	SD	CV	Mean	Mode	SD	CV
Proper treatment	15.47	15.41	0.51	0.03	3.17	3.05	0.46	0.14
Oracle scenario	15.16	15.13	0.47	0.03	3.26	3.15	0.39	0.12
Ignorance scenario	15.65	15.60	0.52	0.03	3.61	3.51	0.43	0.12

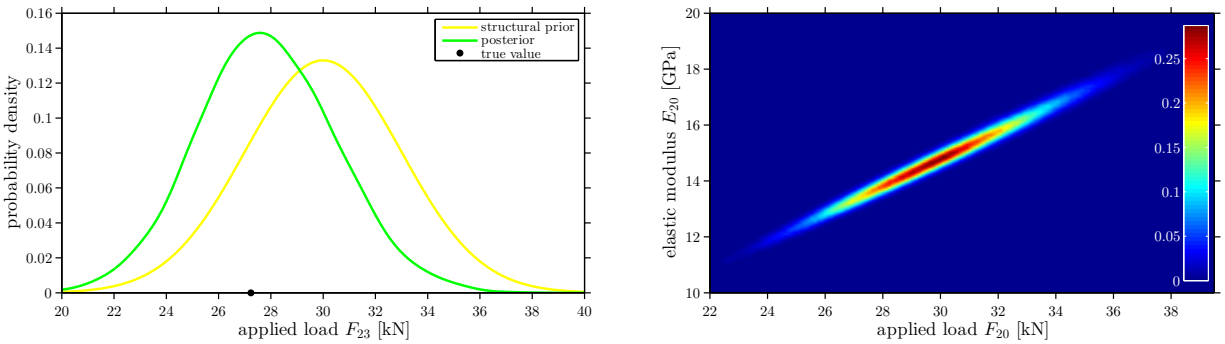
We try to assess the impact of the uncertainty that had been introduced in the loads F_i on the estimation of the QoI $\boldsymbol{\theta}_E = (\mu_E, \sigma_E)$. To that end we pursue the following two strategies. First of all we estimate the QoI while treating the realized loads F_i as if they were part of the experiment-specific knowns $\boldsymbol{d}_i$. This "what-if" or "oracle" scenario actually describes the hypothetical situation that we met in plain probabilistic inversion.

It does not describe the realistic scenario of uncertain conditions $\boldsymbol{\zeta}_i$ that we are actually investigating. Yet this way of proceeding sheds light on how the prescribed uncertainty in the loads affects the inference of the QoI. For $N = 10^7$ and $t = 4.33\,\mathrm{h}$ the results to probabilistic inversion are added to Fig. 4.10. With respect to this idealized situation, one can reassess the previous results of properly treating the loads as uncertain. The introduction of the uncertainty in the loads had actually shifted the posterior modes and raised the level of estimation uncertainty accordingly.

Second of all we investigate the case that the uncertainty $\mathcal{N}(F_i; \mu_{F_i}, \sigma^2_{F_i})$ in the applied loads F_i is simply disregarded. Either it has not been recognized by mistake or it has been intentionally dropped by making simplifying assumptions in favor of a parsimonious model. Rather than treating the loads as belonging to the unknowns $\boldsymbol{\zeta}_i$, we erroneously treat them as such experimental conditions $\boldsymbol{d}_i^{\approx}$ that only approximately describe the prevailing conditions $\boldsymbol{d}_i$. While the data has been created under $\boldsymbol{d}_i$, data analysis is carried out under $\boldsymbol{d}_i^{\approx}$. This describes a situation where the experimenter targets a load $F_i^{\approx} = \mu_{F_i}$, but the testing machine actually realizes F_i. If this uncertainty $\mathcal{N}(F_i; \mu_{F_i}, \sigma^2_{F_i})$ is not accounted for or not recognized at all, the analyst will accomplish inference under the spurious assumption that the loads had taken on their targeted values $F_i^{\approx}$ during experiment execution. For $N = 10^7$ and $t = 3.75\,\mathrm{h}$ the resulting posteriors are added to Fig. 4.10. Our interpretation is that dropping the uncertainty of F_i corrupts the estimation of the QoI and results in misleading estimates of posterior uncertainty, whereas the proper treatment of all uncertainties yields results that are closer to the idealized "oracle" scenario.

4.8.4.2 Results: Intermediate variables

Sampling the joint posterior $\pi(\langle E_i \rangle, \langle F_i \rangle, \boldsymbol{\theta}_E \,|\, \langle \boldsymbol{v}_i \rangle)$ of the entirety of unknowns provides further interesting insights. Apart from the QoI-marginals one can examine the posterior model of experiment-specific loads F_i, notwithstanding that they are considered nuisance. Fig. 4.11 contains two different posteriors involving some F_i. In Fig. 4.11(a) the posterior marginal of a pinpoint load F_i is shown for $i = 23$. The identification of specifically applied loads F_i is subject to rather high levels of posterior uncertainty. This is an issue of statistical identifiability. When both E_i and F_i are uncertain and various combinations of these can explain the observation $\boldsymbol{v}_i$ equally well, then those combinations (E_i, F_i) cannot be distinguished a posteriori. Of course, the reason is that only the ratio F_i/E_i in Eq. (4.39) can be identified. It is therefore interesting to investigate the posterior correlation between the load F_i and the modulus E_i of an experiment i. The two-dimensional posterior of (E_i, F_i) for $i = 20$ that is shown in Fig. 4.11(b) serves as an example. Posterior mass is assigned to suitable parameter constellations (E_i, F_i) that well-explain the measurement $\boldsymbol{v}_i$. As expected the posterior is strongly correlated with a linear coefficient of correlation $r_{F_{20},E_{20}} = 0.99$.



(a) Posterior marginal of F_{23}.

(b) 2D posterior of (F_{20}, E_{20}).

Figure 4.11: Posteriors of intermediate variables. In (a) the posterior marginal of F_{23} and its structural prior $\mathcal{N}(F_{23}; \mu_{F_{23}}, \sigma^2_{F_{23}})$ with $\mu_{F_{23}} = 30\,\mathrm{kN}$ and $\sigma_{F_{23}} = 3\,\mathrm{kN}$ are shown. The posterior is centered around the actual value $F_{23} = 27.24\,\mathrm{kN}$. The two-dimensional posterior of (F_{20}, E_{20}) with $r_{F_{20},E_{20}} = 0.99$ is shown in (b).

4.8.5 Borrowing strength

As pointed out in Section 4.6, Bayesian multilevel modeling allows for "optimal combination of information" or "borrowing strength". Here we demonstrate this inferential mechanism and investigate its underlying flow of information for the previous application example. The Bayesian model of probabilistic inversion Eq. (4.31) is considered. However, as opposed to probabilistic inversion we declare experiment-specific elastic moduli $\langle E_i \rangle$ as

the QoI whereas the hyperparameters $\boldsymbol{\theta}_E$ are considered nuisance. Herein we highlight the optimal inference of a single E_{i_0} for some $i_0 \in \{1, \dots, n\}$.

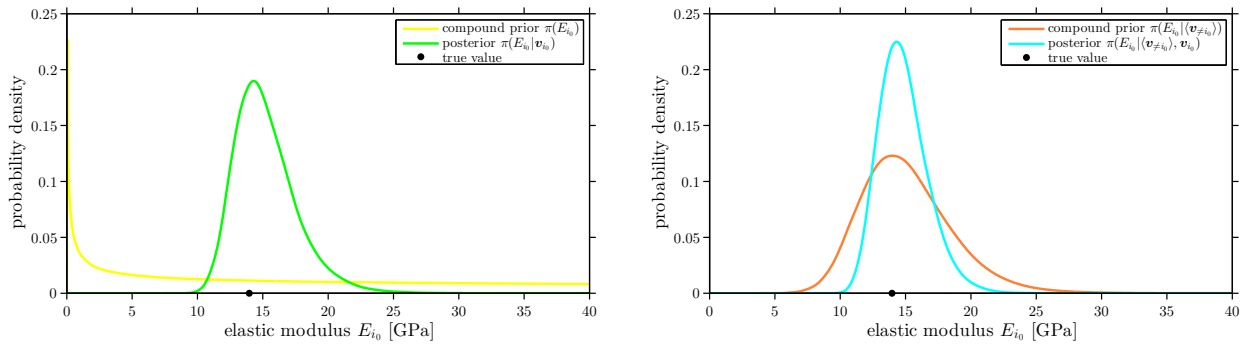
The experimental setup is similar to the one described in Section 4.8.2. For $n = 50$ beams, elastic moduli E_i are randomly sampled from $\mathcal{LN}(E_i|\mu_E, \sigma_E)$. Beam dimensions $\boldsymbol{l}_i$, measurement positions $\boldsymbol{s}_i$ and the applied loads F_i are chosen as before. With Eq. (4.39) beam deflections $\tilde{\boldsymbol{v}}_i$ are predicted. Synthetic data $\boldsymbol{v}_i = \tilde{\boldsymbol{v}}_i + \boldsymbol{\varepsilon}_i$ are generated by perturbing the predictions $\tilde{\boldsymbol{v}}_i$ with noise. For this purpose noise terms $\boldsymbol{\varepsilon}_i$ are independently sampled from Gaussian distributions $\mathcal{N}(\boldsymbol{\varepsilon}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_i)$. We choose $\boldsymbol{\Sigma}_i = \sigma_i^2 \boldsymbol{I}_3$ with $\sigma_i = 0.1\,\text{mm}$ for $i \neq i_0$ and $\sigma_{i_0} = 0.1\,\text{cm}$. The latter describes a comparably large deviation that differs from the setup of Section 4.8.2. This choice serves the purpose of clearly illustrating the inferential mechanism of optimal combination of information.

Eventually optimal combination of information reads as the following problem. With noisy data $\langle \boldsymbol{y}_i \rangle \equiv \langle \boldsymbol{v}_i \rangle$ an experiment-specific $\boldsymbol{x}_{i_0} \equiv E_{i_0}$ has to be ideally estimated, i.e. taking all available sources of information into account. The hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}} \equiv \boldsymbol{\theta}_E$ as well as $\langle \boldsymbol{x}_{\neq i_0} \rangle \equiv \langle E_{\neq i_0} \rangle$ are considered nuisance to that end. Experiment-specific knowns are $\langle \boldsymbol{d}_i \rangle \equiv \langle (F_i, \boldsymbol{l}_i, \boldsymbol{s}_i) \rangle$ and $\langle \boldsymbol{\Sigma}_i \rangle$. The resultant posterior will be of the form $\pi(\boldsymbol{x}_{i_0} | \langle \boldsymbol{y}_i \rangle) \equiv \pi(E_{i_0} | \langle \boldsymbol{v}_i \rangle)$. Subsequent to formulating the joint posterior $\pi(\langle \boldsymbol{x}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle) \equiv \pi(\langle E_i \rangle, \boldsymbol{\theta}_E | \langle \boldsymbol{v}_i \rangle)$, the QoI-marginals can be easily extracted. Other than that, the experimental setup of probabilistic inversion is adopted. Thus the experiment can be visualized by the DAG in Fig. 4.2(b), too.

4.8.5.1 Results: Information accumulation

We conduct simple updating, sequential filtering and multilevel inversion for estimating E_{i_0}, as introduced in Section 4.6. First of all we start with the simple Bayesian updating approach that was introduced in Section 4.6.1. By the method of composition we draw $K = 10^5$ samples $(E_{i_0}^{(1)}, \dots, E_{i_0}^{(K)})$ from the mixture prior $\pi(E_{i_0})$ that corresponds to Eq. (4.35). With this sample the mixture prior can be evaluated as the corresponding one-dimensional KDE with Gaussian kernel functions. The posterior $\pi(E_{i_0} | \boldsymbol{v}_{i_0})$ results from conditioning on the piece of data $\boldsymbol{v}_{i_0}$. This univariate posterior is explored in $N = 10^5$ MCMC iterations for which the program execution time amounts to $t = 5.86\,\text{h}$. The final result of this simple updating approach is shown in Fig. 4.12(a).

Second of all we conduct the sequential Bayesian filtering program that was proposed in Section 4.6.2. In $N = 10^7$ MCMC iterations that take $t = 3.95\,\text{h}$, probabilistic inversion for estimating $\boldsymbol{\theta}_E$ is executed with the data $\langle \boldsymbol{v}_{\neq i_0} \rangle$. MCMC samples from the resultant posterior $\pi(\boldsymbol{\theta}_E | \langle \boldsymbol{v}_{\neq i_0} \rangle)$ are used to sample the compound distribution $\pi(E_{i_0} | \langle \boldsymbol{v}_{\neq i_0} \rangle)$ in Eq. (4.36) via the composition method. Subsequently a lognormal fit to these samples acts as the prior for E_{i_0}. This prior and the arising posterior distribution $\pi(E_{i_0} | \langle \boldsymbol{v}_{\neq i_0} \rangle, \boldsymbol{v}_{i_0})$ are plotted in Fig. 4.12(b). In $t = 0.01\,\text{h}$ of execution time $N = 10^5$ MCMC samples of the univariate posterior were sampled. By comparison of the two posteriors in Fig. 4.12, the shrinkage of the posterior uncertainty from $\pi(E_{i_0} | \boldsymbol{v}_{i_0})$ to $\pi(E_{i_0} | \langle \boldsymbol{v}_{\neq i_0} \rangle, \boldsymbol{v}_{i_0})$ becomes apparent. Both posteriors follow from conditioning on the data $\boldsymbol{v}_{i_0}$, they update different priors $\pi(E_{i_0})$ and $\pi(E_{i_0} | \langle \boldsymbol{v}_{\neq i_0} \rangle)$, though. In the first place this proves that Bayesian priors are a valid source of information. Moreover, this principally shows how learning about E_{i_0} can be indirectly supported by the evidence that $\langle \boldsymbol{v}_{\neq i_0} \rangle$ contains with regard to $\boldsymbol{\theta}_E$.



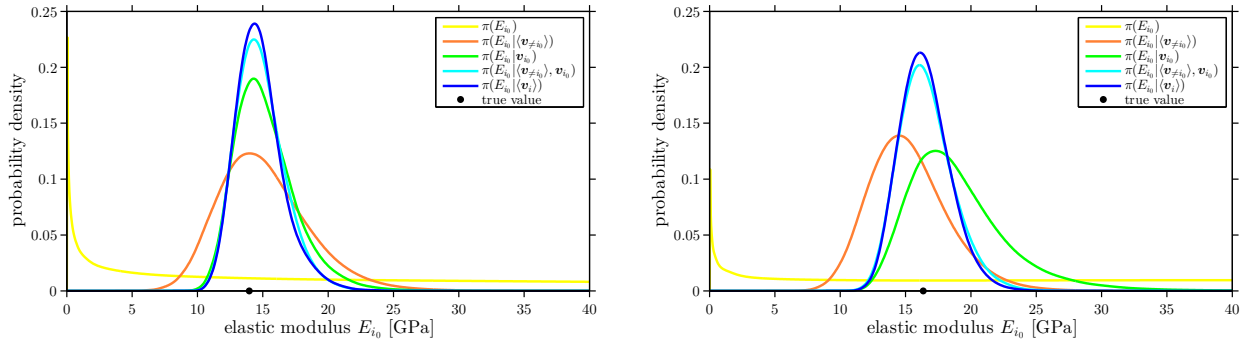
(a) Simple updating.

(b) Sequential filtering.

Figure 4.12: Bayesian updating and filtering. The mixture prior $\pi(E_{i_0})$ and the posterior $\pi(E_{i_0} | \boldsymbol{v}_{i_0})$ of simple updating are shown in (a). Sequential filtering is based on the more informative mixture prior $\pi(E_{i_0} | \langle \boldsymbol{v}_{\neq i_0} \rangle)$ and the corresponding posterior $\pi(E_{i_0} | \langle \boldsymbol{v}_{\neq i_0} \rangle, \boldsymbol{v}_{i_0})$ that are given in (b).

Lastly we perform Bayesian multilevel analysis as described in Section 4.6.3. Sampling the joint posterior $\pi(\langle E_i \rangle, \boldsymbol{\theta}_E | \langle \boldsymbol{v}_i \rangle)$ allows to straightforwardly extract samples from its marginal $\pi(E_{i_0} | \langle \boldsymbol{v}_i \rangle)$ in Eq. (4.37). This is accomplished in $t = 4.57\,\text{h}$ for $N = 10^7$ algorithm iterations. The posterior and the previous inferential distributions relevant for E_{i_0} are plotted in Fig. 4.13. In addition to that Table 4.4 recapitulates the different

approaches. Results are also provided from a second series of runs that were independently carried out on top of the first one. The motivation is to show that borrowing strength is a not a random but a systematic effect. The accumulation of information concerning E_{i_0} manifests in the progressively decreasing uncertainty in the distributions. At every stage of the estimation plan, a certain proportion of the available information has entered the analysis and has been translated into a gain of knowledge related to E_{i_0}. Only the multilevel posterior $\pi(E_{i_0}|\langle \boldsymbol{v}_i \rangle)$ entirely aggregates the available information.



(a) Summary of the 1st series.

(b) Summary of the 2nd series.

Figure 4.13: Accumulation of information. In (a) and (b) the estimations of E_{i_0} are summarized for two series of runs. The true values are $E_{i_0} = 13.96\,\text{GPa}$ and $E_{i_0} = 16.35\,\text{GPa}$ in the 1st and 2nd series, respectively. Uncertainties in identifying these values reflect the amount of information processed in simple updating, sequential filtering and multilevel inversion.

Table 4.4: Posterior summaries of estimating E_{i_0}.

	1st series: E_{i_0} [GPa]			[ ]	2nd series: E_{i_0} [GPa]			[ ]
	Mean	Mode	SD	CV	Mean	Mode	SD	CV
Simple updating	15.23	14.31	2.38	0.16	19.02	17.30	3.93	0.21
Sequential filtering	14.82	14.32	1.83	0.12	16.58	16.07	2.03	0.12
Multilevel inversion	14.75	14.37	1.79	0.12	16.47	16.12	1.85	0.11

The assumption of well-known loads F_i may be overly optimistic in experimental practice. As done in Section 4.8.4 one could attach an additional prescribed uncertainty to those model inputs. In doing so we expect similar results accompanied by a weakening of borrowing strength. Furthermore we expect an indirect form of borrowing strength also to occur for the inputs of a prescribed uncertainty type. Actually the prescribed uncertainty model does not permit for learning about a specific F_{i_0} by borrowing strength directly from $\langle \boldsymbol{v}_{\neq i_0} \rangle$. However, by optimally estimating E_{i_0} also learning F_{i_0} would be indirectly strengthened.

4.9 Conclusion and outlook

Bayesian multilevel model calibration has been developed as a consistent and comprehensive framework for managing uncertainties in inverse problems. At the core of the such problems a forward model relates physical parameters to observable quantities. This deterministic model has been surrounded by a probabilistic representation of uncertainty, variability and error. For this purpose classical Bayesian inversion, hierarchical statistical models and the predominant epistemic/aleatory conception of uncertainty have been utilized. The inferential rationale of multilevel inversion, based on the conditioning, marginalization and transformation of probability measures, has become transparent by laying the research focus on aspects of uncertainty quantification and information accumulation. Fully Bayesian probabilistic inversion and borrowing strength have been suggested. Furthermore we have originally elaborated on the "perfect" data limit. Our developments were driven by the challenges of engineering applications and they ultimately allow for optimal data analysis in intricate situations where evidence is scarce and uncertainty prevails.

An ensemble of structural elements of the same type, for all of which virtual tests are performed and pseudo data are gathered, served as the basis for investigating a variety of experimental scenarios. The amenities of Bayesian multilevel inversion were demonstrated by exercising inference in the chosen example applications

under realistic uncertainty configurations. Probabilistic inversion, i.e. the identification of material variability throughout a population of specimens, was accomplished and it was investigated how the amount of data influences the estimation uncertainty . The constraints of perfectly known residual variances and experimental conditions were loosened. In this context we calibrated the forward model prediction error and we studied how the objective of probabilistic inversion is impeded by additional uncertainties in the experimental conditions. Optimal combination of information, i.e. the ideal inference of specimen-specific properties, has been introduced as a byproduct of the joint formulation of multilevel inversion. Especially in the engineering community this is an aspect that is often overlooked. We examined the underlying inferential mechanisms and we identified the computational obstacles, e.g. costly evaluations of the marginalized likelihood function or the curse of high-dimensionality.

In conclusion, innovative techniques must be developed in order to overcome these difficulties for solving "real-world" problems. Future research therefore includes the following items. For the marginal problem, numerically efficient and acceptably accurate approximations of the integrated likelihood have to be developed. Advanced MCMC techniques, that are custom-tailored for the specific structure of multilevel posteriors, have to be devised for the joint problem. In this connection a numerical study involving HMC is in progress. For both the marginal and the joint variant of multilevel inversion, the application of dedicated metamodeling techniques promises drastic speedups. It will also be interesting to study the applicability and performance of optimal transportation approaches [93, 94] to classical Bayesian inference in the context of multilevel estimation. Another research question concerns the role of multimodality and severe ill-posedness of separate inverse problems in Bayesian multilevel inversion.

References

[1] A. M. Stuart. "Inverse problems: A Bayesian perspective". In: *Acta Numerica* 19 (2010), pp. 451–559. DOI: 10.1017/S0962492910000061.

[2] M. Allmaras, W. Bangerth, J. Linhart, J. Polanco, F. Wang, K. Wang, J. Webster, and S. Zedler. "Estimating Parameters in Physical Models through Bayesian Inversion: A Complete Example". In: *SIAM Review* 55.1 (2013), pp. 149–167. DOI: 10.1137/100788604.

[3] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2005. DOI: 10.1137/1.9780898717921.

[4] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Applied Mathematical Sciences 160. New York: Springer, 2005. DOI: 10.1007/b138659.

[5] R. Hadidi and N. Gucunski. "Probabilistic Approach to the Solution of Inverse Problems in Civil Engineering". In: *Journal of Computing in Civil Engineering* 22.6 (2008), pp. 338–347. DOI: 10.1061/(ASCE)0887-3801(2008)22:6(338).

[6] J. L. Beck. "Bayesian system identification based on probability logic". In: *Structural Control and Health Monitoring* 17.7 (2010), pp. 825–847. DOI: 10.1002/stc.424.

[7] A. Malinverno and V. A. Briggs. "Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes". In: *Geophysics* 69.4 (2004), pp. 1005–1016. DOI: 10.1190/1.1778243.

[8] J. Wang and N. Zabaras. "Hierarchical Bayesian models for inverse problems in heat conduction". In: *Inverse Problems* 21.1 (2005), pp. 183–206. DOI: 10.1088/0266-5611/21/1/012.

[9] L. Wu. *Mixed Effects Models for Complex Data*. Monographs on Statistics & Applied Probability 113. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2010. DOI: 10.1201/9781420074086.

[10] S. W. Raudenbush. "Educational Applications of Hierarchical Linear Models: A Review". In: *Journal of Educational Statistics* 13.2 (1988), pp. 85–116. DOI: 10.3102/10769986013002085.

[11] M. H. Seltzer, W. H. Wong, and A. S. Bryk. "Bayesian Analysis in Applications of Hierarchical Models: Issues and Methods". In: *Journal of Educational and Behavioral Statistics* 21.2 (1996), pp. 131–167. DOI: 10.3102/10769986021002131.

[12] J. Wakefield. "The Bayesian Analysis of Population Pharmacokinetic Models". In: *Journal of the American Statistical Association* 91.433 (1996), pp. 62–75. DOI: 10.1080/01621459.1996.10476664.

[13] H. T. Banks and L. K. Potter. "Probabilistic methods for addressing uncertainty and variability in biological models: application to a toxicokinetic model". In: *Mathematical Biosciences* 192.2 (2004), pp. 193–225. DOI: 10.1016/j.mbs.2004.11.008.

[14] M. Davidian and D. M. Giltinan. "Nonlinear Models for Repeated Measurement Data: An Overview and Update". In: *Journal of Agricultural, Biological, and Environmental Statistics* 8.4 (2003), pp. 387–419. DOI: 10.1198/1085711032697.

[15] H. T. Banks, Z. R. Kenz, and W. C. Thompson. "A review of selected techniques in inverse problem nonparametric probability distribution estimation". In: *Journal of Inverse and Ill-posed Problems* 20.4 (2012), pp. 429–460. DOI: 10.1515/jip-2012-0037.

[16] M. Davidian and D. M. Giltinan. *Nonlinear Models for Repeated Measurement Data*. Monographs on Statistics & Applied Probability 62. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1995.

[17] H. T. Banks, S. Hu, and W. C. Thompson. *Modeling and Inverse Problems in the Presence of Uncertainty*. Monographs and Research Notes in Mathematics. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2014.

[18] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press, 2006.

[19] P. D. Congdon. *Applied Bayesian Hierarchical Methods*. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2010. DOI: 10.1201/9781584887218.

[20] S. Jackman. *Bayesian Analysis for the Social Sciences*. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2009. DOI: 10.1002/9780470686621.

[21] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. 3rd ed. Texts in Statistical Science. Boca Raton, Florida, USA: CRC Press, 2014.

[22] M. H. Faber. "On the Treatment of Uncertainties and Probabilities in Engineering Decision Analysis". In: *Journal of Offshore Mechanics and Arctic Engineering* 127.3 (2005), pp. 243–248. DOI: 10.1115/1.1951776.

[23] A. Der Kiureghian and O. Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009), pp. 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

[24] J. C. Helton and W. L. Oberkampf. "Alternative representations of epistemic uncertainty". In: *Reliability Engineering & System Safety* 85.1–3 (2004), pp. 1–10. DOI: 10.1016/j.ress.2004.03.001.

[25] J. C. Helton and J. D. Johnson. "Quantification of margins and uncertainties: Alternative representations of epistemic uncertainty". In: *Reliability Engineering & System Safety* 96.9 (2011), pp. 1034–1052. DOI: 10.1016/j.ress.2011.02.013.

[26] T. Koski and J. M. Noble. *Bayesian Networks: An Introduction*. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2009. DOI: 10.1002/9780470684023.

[27] U. B. Kjærulff and A. L. Madsen. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. 2nd ed. Information Science and Statistics 22. New York: Springer, 2013. DOI: 10.1007/978-0-387-74101-7.

[28] Y. Y. Bayraktarli, J. W. Baker, and M. H. Faber. "Uncertainty treatment in earthquake modelling using Bayesian probabilistic networks". In: *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 5.1 (2011), pp. 44–58. DOI: 10.1080/17499511003679931.

[29] M. Deublein, M. Schubert, B. T. Adey, J. Köhler, and M. H. Faber. "Prediction of road accidents: A Bayesian hierarchical approach". In: *Accident Analysis & Prevention* 51 (2013), pp. 274–291. DOI: 10.1016/j.aap.2012.11.019.

[30] D. L. Kelly and C. L. Smith. "Bayesian inference in probabilistic risk assessment – The current state of the art". In: *Reliability Engineering & System Safety* 94.2 (2009), pp. 628–643. DOI: 10.1016/j.ress.2008.07.002.

[31] A. Urbina, S. Mahadevan, and T. L. Paez. "A Bayes Network Approach to Uncertainty Quantification in Hierarchically Developed Computational Models". In: *International Journal for Uncertainty Quantification* 2.2 (2012), pp. 173–193. DOI: 10.1615/Int.J.UncertaintyQuantification.v2.i2.70.

[32] E. de Rocquigny and S. Cambier. "Inverse probabilistic modelling of the sources of uncertainty: A non-parametric simulated-likelihood method with application to an industrial turbine vibration assessment". In: *Inverse Problems in Science and Engineering* 17.7 (2009), pp. 937–959. DOI: 10.1080/17415970902916987.

[33] G. Celeux, A. Grimaud, Y. Lefèbvre, and E. de Rocquigny. "Identifying intrinsic variability in multivariate systems through linearized inverse methods". In: *Inverse Problems in Science and Engineering* 18.3 (2010), pp. 401–415. DOI: 10.1080/17415971003624330.

[34] P. Barbillon, G. Celeux, A. Grimaud, Y. Lefèbvre, and E. de Rocquigny. "Nonlinear methods for inverse statistical problems". In: *Computational Statistics & Data Analysis* 55.1 (2011), pp. 132–142. DOI: 10.1016/j.csda.2010.05.030.

[35] E. de Rocquigny. *Modelling Under Risk and Uncertainty: An Introduction to Statistical, Phenomenological and Computational Methods*. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2012. DOI: 10.1002/9781119969495.

[36] J. O. Berger and W. E. Strawderman. "Choice of hierarchical priors: admissibility in estimation of normal means". In: *The Annals of Statistics* 24.3 (1996), pp. 931–951. DOI: 10.1214/aos/1032526950.

[37] J. O. Berger, W. Strawderman, and D. Tang. "Posterior Propriety and Admissibility of Hyperpriors in Normal Hierarchical Models". In: *The Annals of Statistics* 33.2 (2005), pp. 606–646. DOI: 10.1214/009053605000000075.

[38] A. Gelman. "Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper)". In: *Bayesian Analysis* 1.3 (2006), pp. 515–534. DOI: 10.1214/06-BA117A.

[39] D. Draper, J. S. Hodges, C. L. Mallows, and D. Pregibon. "Exchangeability and Data Analysis". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 156.1 (1993), pp. 9–37. DOI: 10.2307/2982858.

[40] J. M. Bernardo. "The Concept of Exchangeability and its Applications". In: *Far East Journal of Mathematical Sciences* 4 (1996), pp. 111–121.

[41] E. Simoen, C. Papadimitriou, and G. Lombaert. "On prediction error correlation in Bayesian model updating". In: *Journal of Sound and Vibration* 332.18 (2013), pp. 4136–4152. DOI: 10.1016/j.jsv.2013.03.019.

[42] E. L. Zhang, P. Feissel, and J. Antoni. "A comprehensive Bayesian approach for model updating and quantification of modeling errors". In: *Probabilistic Engineering Mechanics* 26.4 (2011), pp. 550–560. DOI: 10.1016/j.probengmech.2011.07.001.

[43] M. C. Kennedy and A. O'Hagan. "Bayesian calibration of computer models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 425–464. DOI: 10.1111/1467-9868.00294.

[44] P. D. Arendt, D. W. Apley, and W. Chen. "Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability". In: *Journal of Mechanical Design* 134.10, 100908 (2012), pp. 1–12. DOI: 10.1115/1.4007390.

[45] E. L. Droguett and A. Mosleh. "Bayesian Methodology for Model Uncertainty Using Model Performance Data". In: *Risk Analysis* 28.5 (2008), pp. 1457–1476. DOI: 10.1111/j.1539-6924.2008.01117.x.

[46] I. Park and R. V. Grandhi. "A Bayesian statistical method for quantifying model form uncertainty and two model combination methods". In: *Reliability Engineering & System Safety* 129 (2014), pp. 46–56. DOI: 10.1016/j.ress.2014.04.023.

[47] J. Beck and K. Yuen. "Model Selection Using Response Measurements: Bayesian Probabilistic Approach". In: *Journal of Engineering Mechanics* 130.2 (2004), pp. 192–203. DOI: 10.1061/(ASCE)0733-9399(2004)130:2(192).

[48] K.-V. Yuen. "Recent developments of Bayesian model class selection and applications in civil engineering". In: *Structural Safety* 32.5 (2010), pp. 338–346. DOI: 10.1016/j.strusafe.2010.03.011.

[49] D. Draper. "Assessment and Propagation of Model Uncertainty". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 45–97.

[50] D. Basu. "On the Elimination of Nuisance Parameters". In: *Journal of the American Statistical Association* 72.358 (1977), pp. 355–366. DOI: 10.1080/01621459.1977.10481002.

[51] A. P. Dawid. "A Bayesian Look at Nuisance Parameters". In: *Trabajos de Estadística Y de Investigación Operativa* 31.1 (1980), pp. 167–203. DOI: 10.1007/BF02888351.

[52] J. O. Berger, B. Liseo, and R. L. Wolpert. "Integrated Likelihood Methods for Eliminating Nuisance Parameters". In: *Statistical Science* 14.1 (1999), pp. 1–28. DOI: 10.1214/ss/1009211804.

[53] T. A. Severini. "On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters". In: *Statistica Sinica* 9.3 (1999), pp. 713–724.

[54] T. A. Severini. "Integrated likelihood functions for non-Bayesian inference". In: *Biometrika* 94.3 (2007), pp. 529–542. DOI: 10.1093/biomet/asm040.

[55] M. A. Beaumont. "Estimation of Population Growth or Decline in Genetically Monitored Populations". In: *Genetics* 164.3 (2003), pp. 1139–1160.

[56] Y. J. Sung and C. J. Geyer. "Monte Carlo Likelihood Inference for Missing Data Models". In: *The Annals of Statistics* 35.3 (2007), pp. 990–1011. DOI: 10.1214/009053606000001389.

[57] C. S. Bos. "A Comparison of Marginal Likelihood Computation Methods". In: *Compstat: Proceedings in Computational Statistics*. Ed. by W. Härdle and B. Rönz. Physica-Verlag Heidelberg New York, 2002, pp. 111–116. DOI: 10.1007/978-3-642-57489-4_11.

[58] L. G. Crespo, S. P. Kenny, and D. P. Giesy. "The NASA Langley Multidisciplinary Uncertainty Quantification Challenge". In: *16th AIAA Non-Deterministic Approaches Conference (SciTech 2014)*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2014. DOI: 10.2514/6.2014-1347.

[59] J. B. Nagel and B. Sudret. "A Bayesian Multilevel Framework for Uncertainty Characterization and the NASA Langley Multidisciplinary UQ Challenge". In: *16th AIAA Non-Deterministic Approaches Conference (SciTech 2014)*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2014. DOI: 10.2514/6.2014-1502.

[60] J. B. Nagel and B. Sudret. "Bayesian Multilevel Model Calibration for Inverse Problems Under Uncertainty with Perfect Data". In: *Journal of Aerospace Information Systems* 12.1 (2015), pp. 97–113. DOI: 10.2514/1.I010264.

[61] B. Sudret, F. Perrin, and M. Pendola. "Use of polynomial chaos expansions in stochastic inverse problems". In: *4th International ASRANet Colloquium*. Glasgow, Scotland, UK: ASRANet Ltd, 2008.

[62] Y. M. Marzouk, H. N. Najm, and L. A. Rahn. "Stochastic spectral methods for efficient Bayesian solution of inverse problems". In: *Journal of Computational Physics* 224.2 (2007), pp. 560–586. DOI: 10.1016/j.jcp.2006.10.010.

[63] Y. M. Marzouk and H. N. Najm. "Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems". In: *Journal of Computational Physics* 228.6 (2009), pp. 1862–1902. DOI: 10.1016/j.jcp.2008.11.024.

[64] P. M. Tagade and H.-L. Choi. "A Generalized Polynomial Chaos-Based Method for Efficient Bayesian Calibration of Uncertain Computational Models". In: *Inverse Problems in Science and Engineering* 22.4 (2014), pp. 602–624. DOI: 10.1080/17415977.2013.823411.

[65] J. Beck and S. Au. "Bayesian Updating of Structural Models and Reliability using Markov Chain Monte Carlo Simulation". In: *Journal of Engineering Mechanics* 128.4 (2002), pp. 380–391. DOI: 10.1061/(ASCE)0733-9399(2002)128:4(380).

[66] J. Ching and Y. Chen. "Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection, and Model Averaging". In: *Journal of Engineering Mechanics* 133.7 (2007), pp. 816–832. DOI: 10.1061/(ASCE)0733-9399(2007)133:7(816).

[67] C. Papadimitriou and D.-C. Papadioti. "Component mode synthesis techniques for finite element model updating". In: *Computers & Structures* 126 (2013), pp. 15–28. DOI: 10.1016/j.compstruc.2012.10.018.

[68] H. A. Jensen, E. Millas, D. Kusanovic, and C. Papadimitriou. "Model-reduction techniques for Bayesian finite element model updating using dynamic response data". In: *Computer Methods in Applied Mechanics and Engineering* 279 (2014), pp. 301–324. DOI: 10.1016/j.cma.2014.06.032.

[69] J. B. Nagel and B. Sudret. "Probabilistic Inversion for Estimating the Variability of Material Properties: A Bayesian Multilevel Approach". In: *11th International Probabilistic Workshop (IPW11)*. Ed. by D. Novák and M. Vořechovský. Brno, Czech Republic: Litera, 2013, pp. 293–303. DOI: 10.3929/ethz-a-010034843.

[70] G. C. Ballesteros, P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos. "Bayesian Hierarchical Models for Uncertainty Quantification in Structural Dynamics". In: *2nd International Conference on Vulnerability and Risk Analysis and Management and 6th International Symposium on Uncertainty Modeling and Analysis (ICVRAM & ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 162, pp. 1615–1624. DOI: 10.1061/9780784413609.162.

[71] B. Kraan and T. Bedford. "Probabilistic Inversion of Expert Judgments in the Quantification of Model Uncertainty". In: *Management Science* 51.6 (2005), pp. 995–1006. DOI: 10.1287/mnsc.1050.0370.

[72] C. Du, D. Kurowicka, and R. M. Cooke. "Techniques for generic probabilistic inversion". In: *Computational Statistics & Data Analysis* 50.5 (2006), pp. 1164–1187. DOI: 10.1016/j.csda.2005.01.002.

[73] C. Desceliers, R. Ghanem, and C. Soize. "Maximum likelihood estimation of stochastic chaos representations from experimental data". In: *International Journal for Numerical Methods in Engineering* 66.6 (2006), pp. 978–1001. DOI: 10.1002/nme.1576.

[74] C. Desceliers, C. Soize, and R. Ghanem. "Identification of Chaos Representations of Elastic Properties of Random Media Using Experimental Vibration Tests". In: *Computational Mechanics* 39.6 (2007), pp. 831–838. DOI: 10.1007/s00466-006-0072-7.

[75] L. Mehrez, A. Doostan, D. Moens, and D. Vandepitte. "Stochastic identification of composite material properties from limited experimental databases, Part II: Uncertainty modelling". In: *Mechanical Systems and Signal Processing* 27 (2012), pp. 484–498. DOI: 10.1016/j.ymssp.2011.09.001.

[76] S. Debruyne, D. Vandepitte, and D. Moens. "Identification of design parameter variability of honeycomb sandwich beams from a study of limited available experimental dynamic structural response data". In: *Computers & Structures* 146 (2015), pp. 197–213. DOI: 10.1016/j.compstruc.2013.09.004.

[77] J. B. Nagel and B. Sudret. "A Bayesian Multilevel Approach to Optimally Estimate Material Properties". In: *2nd International Conference on Vulnerability and Risk Analysis and Management and 6th International Symposium on Uncertainty Modeling and Analysis (ICVRAM & ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 151, pp. 1504–1513. DOI: 10.1061/9780784413609.151.

[78] D. Draper, D. P. Gaver, P. K. Goel, J. B. Greenhouse, L. V. Hedges, C. N. Morris, and C. M. Waternaux. *Combining Information: Statistical Issues and Opportunities for Research*. Washington, D.C., USA: Panel on Statistical Issues et al., 1992.

[79] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1996.

[80] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. 2nd ed. Springer Series in Statistics. New York: Springer, 2004. DOI: 10.1007/978-1-4757-4145-2.

[81] M. K. Cowles and B. P. Carlin. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review". In: *Journal of the American Statistical Association* 91.434 (1996), pp. 883–904. DOI: 10.1080/01621459.1996.10476956.

[82] S. P. Brooks and G. O. Roberts. "Convergence assessment techniques for Markov chain Monte Carlo". In: *Statistics and Computing* 8.4 (1998), pp. 319–335. DOI: 10.1023/A:1008820505350.

[83] A. Gelman and D. B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (1992), pp. 457–472. DOI: 10.1214/ss/1177011136.

[84] S. P. Brooks and A. Gelman. "General Methods for Monitoring Convergence of Iterative Simulations". In: *Journal of Computational and Graphical Statistics* 7.4 (1998), pp. 434–455. DOI: 10.1080/10618600.1998.10474787.

[85] C. J. Geyer. "Introduction to Markov Chain Monte Carlo". In: *Handbook of Markov Chain Monte Carlo*. Ed. by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. Handbooks of Modern Statistical Methods. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2011. Chap. 1, pp. 3–48. DOI: 10.1201/b10905-2.

[86] P. D. O'Neill, D. J. Balding, N. G. Becker, M. Eerola, and D. Mollison. "Analyses of Infectious Disease Data from Household Outbreaks by Markov chain Monte Carlo Methods". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49.4 (2000), pp. 517–542. DOI: 10.1111/1467-9876.00210.

[87] G. Bal, I. Langmore, and Y. Marzouk. "Bayesian Inverse Problems with Monte Carlo Forward Models". In: *Inverse Problems and Imaging* 7.1 (2013), pp. 81–105. DOI: 10.3934/ipi.2013.7.81.

[88] D. A. van Dyk and X.-L. Meng. "The Art of Data Augmentation". In: *Journal of Computational and Graphical Statistics* 10.1 (2001), pp. 1–50. DOI: 10.1198/10618600152418584.

[89] D. A. van Dyk. "Hierarchical Models, Data Augmentation, and Markov Chain Monte Carlo". In: *Statistical Challenges in Astronomy*. New York: Springer, 2003, pp. 41–55. DOI: 10.1007/0-387-21529-8_3.

[90] C. Andrieu and G. O. Roberts. "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations". In: *The Annals of Statistics* 37.2 (2009), pp. 697–725. DOI: 10.1214/07-AOS574.

[91] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (1987), pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.

[92] R. M. Neal. "MCMC Using Hamiltonian Dynamics". In: *Handbook of Markov Chain Monte Carlo*. Ed. by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. Handbooks of Modern Statistical Methods. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2011. Chap. 5, pp. 113–162. DOI: 10.1201/b10905-6.

[93] S. Reich. "A dynamical systems framework for intermittent data assimilation". In: *BIT Numerical Mathematics* 51.1 (2011), pp. 235–249. DOI: 10.1007/s10543-010-0302-4.

[94] T. A. El Moselhy and Y. M. Marzouk. "Bayesian inference with optimal maps". In: *Journal of Computational Physics* 231.23 (2012), pp. 7815–7850. DOI: 10.1016/j.jcp.2012.07.022.

Chapter 5

Hamiltonian Monte Carlo in hierarchical inverse problems

Original publication

J. B. Nagel and B. Sudret. "Hamiltonian Monte Carlo and Borrowing Strength in Hierarchical Inverse Problems". In: *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 2.3, B4015008 (2016), pp. 1–12. DOI: 10.1061/AJRUA6.0000847

Abstract

Bayesian approaches to uncertainty quantification and information acquisition in hierarchically defined inverse problems are presented. The techniques comprise simple updating, staged estimation and multilevel model calibration. In particular the estimation of material properties within an ensemble of identically manufactured structural elements is considered. It is shown how inferring the characteristics of an individual specimen can be accomplished by exhausting statistical strength from tests of other ensemble members. This is useful in experimental situations where evidence is scarce or unequally distributed. Hamiltonian Monte Carlo is proposed in order to cope with the numerical challenges of the devised approaches. The performance of the algorithm is studied and compared to classical Markov chain Monte Carlo sampling. It turns out that Bayesian posterior computations can be drastically accelerated.

5.1 Introduction

Bayesian inference establishes a flexible framework for solving inverse problems. Given measured responses of a forward model, it allows for reducing epistemic uncertainty of unknown parameters [1, 2] and, within the context of hierarchical modeling, for quantifying the aleatory uncertainty of unobservables [3, 4]. The latter type of problem is often encountered in social science or biological statistics. Yet it is of great interest for engineering applications, too. It allows one to study the natural variability of physical parameters that cannot be directly measured. This involves the variability of material properties as a result of the unavoidable uncertainties in the manufacturing process or due to spatial and temporal changes in the environmental conditions [5, 6].

In the recent literature a number of approaches have been devised that aim at fitting the hyperparameters of the aleatory distribution of forward model inputs [7–9]. These approaches build upon the marginalization of varying inputs at the level of the likelihood function. This way of proceeding typically leads to low-dimensional estimation problems where the major difficulty lies in computing the integrated likelihood. The joint inference of the distributional hyperparameters and the experiment-specific realizations of the variables constitutes a higher-dimensional problem [10–12]. Even though the marginalized and the joint problem variants are equivalent with regard to hyperparameter calibration, the former formulation does not allow for the estimation of realizations of the variable parameters. To this effect one has to rely on the joint problem formulation.

In this paper the joint parameter/hyperparameter inference is studied in view of reducing the uncertainty in the parameters. The goal is to demonstrate the advantages of this formulation and to overcome its computational difficulties. First, we investigate borrowing strength [13] as a means of information aggregation in inverse problems. This statistical mechanism allows for an optimal estimation of individual material properties within a specimen ensemble. We prove that this is valuable in experimental situations where uncertainty is dominant and information is heterogeneous. Indeed these are characteristics of problems in civil engineering. Specifically the

system under consideration is an ensemble of identically manufactured beams that are individually tested in a series of experiments. A situation is considered where evidence is unevenly distributed throughout the ensemble members, i.e. the properties of some members can be measured with high accuracy whereas others are poorly informed by the data. Second, in order to alleviate problems of Markov chain Monte Carlo (MCMC) [14, 15] for posterior exploration in high-dimensional parameter spaces, we propose Hamiltonian Monte Carlo (HMC) [16]. The latter is an advanced MCMC technique that allows for gradient-assisted posterior computation with auxiliary variables. We show that HMC is ideally suited and extremely efficient for borrowing strength in inverse problems.

The main part of the paper is structured in the following way. Stochastic inversion and multilevel modeling are reviewed first. On this basis, borrowing strength and information accumulation are investigated. An introduction to HMC sampling is provided after that. In order to study the proposed methods, a numerical experiment with simulated data is conducted. Concluding remarks are given in the end.

5.2 Multilevel inversion

Inversion is the inference of model parameters from noisy and limited data and is often formulated as statistical estimation. This formulation encompasses a wide range of problems. *Parameter estimation* [1, 2] aims at inferring unknown parameters $\boldsymbol{x} \in \mathbb{R}^m$ of a physical forward model $\mathcal{M}$. This model predicts the outcome $\mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_i)$ of $i = 1, \dots, n$ experiments under known experimental conditions $\boldsymbol{d}_i$. The discrepancy between predictions and real data $\boldsymbol{y}_i$ is accounted for by a statistical model $\boldsymbol{y}_i = \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_i) + \boldsymbol{\varepsilon}_i$. Here the residuals $\boldsymbol{\varepsilon}_i \sim f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i)$ capture measurement errors, numerical approximations and model inadequacies. A widespread probabilistic formulation rests on independent Gaussian distributions $f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i) = \mathcal{N}(\boldsymbol{\varepsilon}_i | \boldsymbol{0}, \boldsymbol{\Sigma}_i)$ with covariance matrices $\boldsymbol{\Sigma}_i$. In Bayesian inversion the prior distribution $\pi(\boldsymbol{x})$ quantifies the epistemic parameter uncertainty of the quantities of interest (QoI) $\boldsymbol{x}$ before the data are analyzed. With the likelihood function $\mathcal{L}(\boldsymbol{x}) = \prod_{i=1}^n f_{\boldsymbol{E}_i}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}, \boldsymbol{d}_i))$ the posterior probability density follows through Bayes' law $\pi(\boldsymbol{x} | \langle \boldsymbol{y}_i \rangle) \propto \mathcal{L}(\boldsymbol{x})\, \pi(\boldsymbol{x})$. Here the data from all experiments are denoted as $\langle \boldsymbol{y}_i \rangle = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$. The posterior represents the knowledge about the unknown parameter vector after conditioning on the data. Apart from exceptional cases the posterior has no analytically closed form, thus more often than not the challenge lies in sampling the posterior by means of MCMC [14, 15].

Another type of inverse problem arises when variations in the data $\boldsymbol{y}_i$ are not only attributed to different conditions $\boldsymbol{d}_i$ and random residuals $\boldsymbol{\varepsilon}_i$, but also as a consequence of naturally varying forward model inputs $\boldsymbol{x}_i$. Throughout the experiments $i = 1, \dots, n$ those inputs take on unobservable realizations of conditionally independent random variables $(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X} | \boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})$. This conditional distribution represents aleatory variability. Data are then represented as $\boldsymbol{y}_i = \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i) + \boldsymbol{\varepsilon}_i$. Prior to data analysis the distribution $\pi(\boldsymbol{\theta}_{\boldsymbol{X}})$ encodes the epistemic uncertainty of the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$. The described experimental situation is summarized as the hierarchical model

$$(\boldsymbol{y}_i | \boldsymbol{x}_i) \sim f_{\boldsymbol{E}_i}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i)), \tag{5.1a}$$

$$(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X} | \boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}), \tag{5.1b}$$

$$\boldsymbol{\theta}_{\boldsymbol{X}} \sim \pi(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{5.1c}$$

In Fig. 5.1 this model is visualized as a directed acyclic graph (DAG). Here nodes represent known or unknown quantities while directed edges represent probabilistic or deterministic relations.

Probabilistic inversion [7–9] is the problem of estimating the unknown hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ that are the QoI. A likelihood for this class of problems can be obtained by the marginalization $\mathcal{L}(\boldsymbol{\theta}_{\boldsymbol{X}}) = \prod_{i=1}^n \int f_{\boldsymbol{E}_i}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i))\, f_{\boldsymbol{X} | \boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})\, \mathrm{d}\boldsymbol{x}_i$. The posterior $\pi(\boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle) \propto \mathcal{L}(\boldsymbol{\theta}_{\boldsymbol{X}})\, \pi(\boldsymbol{\theta}_{\boldsymbol{X}})$ results from Bayes' theorem. In practice the integrated likelihood can be computed through stochastic integration [17] or Laplace's method [18]. *Multilevel inversion* [12, 19] is the joint estimation of all unknowns $(\langle \boldsymbol{x}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}})$ by conditioning on all knowns $\langle \boldsymbol{y}_i \rangle$. The corresponding joint posterior distribution is given as

$$\pi(\langle \boldsymbol{x}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle) \propto \left(\prod_{i=1}^n f_{\boldsymbol{E}_i}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i)) \right) \left(\prod_{i=1}^n f_{\boldsymbol{X} | \boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \right) \pi(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{5.2}$$

On the one side, the posterior Eq. (5.2) offers the possibility to pool information. Individual realizations $\boldsymbol{x}_i$ can be optimally inferred. This is known as *optimal combination of information* or simply as *borrowing strength* [13]. In the subsequent section this possibility is investigated. On the downside, the high-dimensionality of the parameter space is a serious challenge that may necessitate advanced MCMC sampling schemes. The number of parameters in the vector determines the dimensionality of the parameter space. Let l and m denote the dimensions of the spaces of the unknowns $\boldsymbol{\theta}_{\boldsymbol{X}}$ and $\boldsymbol{x}_i$, respectively. Then the posterior in Eq. (5.2) involves a

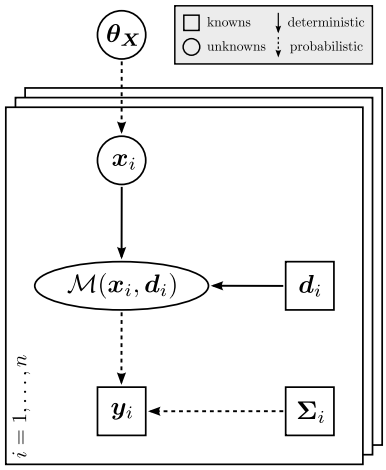


Figure 5.1: DAG of the multilevel model.

$(l + m \cdot n)$-dimensional parameter space, i.e. the dimension grows linearly with the sample size n. Unfortunately this means that the computational cost increases with the number of experiments conducted. In order to ameliorate this situation, HMC is later proposed as an efficient means to explore joint posteriors of the form Eq. (5.2).

5.3 Combination of information

Information-wise the estimation of the unknowns in Eq. (5.1) can be based on the observed data $\boldsymbol{y}_i$, the Bayesian prior $\pi(\boldsymbol{\theta}_{\boldsymbol{X}})$, the structural knowledge $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})$ and the information encapsulated in $f_{\boldsymbol{E}_i}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{d}_i))$. Now we focus on the optimal inference of an individual parameter $\boldsymbol{x}_{i_0}$ for some $i_0 \in \{1, \ldots, n\}$. Instead of merely inverting the observation $\boldsymbol{y}_{i_0}$ for the corresponding $\boldsymbol{x}_{i_0}$, we solve the joint multilevel problem. As it turns out, in doing so one can obtain more information about $\boldsymbol{x}_{i_0}$ than what is contained in $\boldsymbol{y}_{i_0}$. One can indirectly learn from the data $\langle \boldsymbol{y}_{\neq i_0} \rangle = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{i_0-1}, \boldsymbol{y}_{i_0+1}, \ldots, \boldsymbol{y}_n)$ that were collected in different experiments. This is beneficial in the event of that the tests in experiment i_0 are less informative, e.g. the associated data $\boldsymbol{y}_{i_0}$ are less numerous or subject to a higher degree of measurement uncertainty. In order to demonstrate the effect and to understand its underlying information flow, we pursue the following three strategies for the inference of $\boldsymbol{x}_{i_0}$.

5.3.1 Simple updating

In this first approach inference of $\boldsymbol{x}_{i_0}$ is based on the data $\boldsymbol{y}_{i_0}$, the structural knowledge $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_{i_0}|\boldsymbol{\theta}_{\boldsymbol{X}})$ and the prior $\pi(\boldsymbol{\theta}_{\boldsymbol{X}})$. By marginalizing the joint prior $\pi(\boldsymbol{x}_{i_0}, \boldsymbol{\theta}_{\boldsymbol{X}}) = f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_{i_0}|\boldsymbol{\theta}_{\boldsymbol{X}})\,\pi(\boldsymbol{\theta}_{\boldsymbol{X}})$ over the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$, the prior distribution of $\boldsymbol{x}_{i_0}$ is written as

$$\pi(\boldsymbol{x}_{i_0}) = \int f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_{i_0}|\boldsymbol{\theta}_{\boldsymbol{X}})\,\pi(\boldsymbol{\theta}_{\boldsymbol{X}})\,\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}. \tag{5.3}$$

This compound probability distribution represents the uncertainty of $\boldsymbol{x}_{i_0}$ prior to data analysis. Simple updating of the prior $\pi(\boldsymbol{x}_{i_0})$ by conditioning on $\boldsymbol{y}_{i_0}$ leads to the posterior $\pi(\boldsymbol{x}_{i_0}|\boldsymbol{y}_{i_0})$. While the observation $\boldsymbol{y}_{i_0}$ has entered the analysis of $\boldsymbol{x}_{i_0}$, the data $\langle \boldsymbol{y}_{\neq i_0} \rangle$ have been neglected. In other words, the hierarchical problem structure has been recognized but it has not yet been fully utilized. In constructing the prior Eq. (5.3) it has been acknowledged that information about $\boldsymbol{\theta}_{\boldsymbol{X}}$ carries information about $\boldsymbol{x}_{i_0}$. However, the uncertainty in $\boldsymbol{\theta}_{\boldsymbol{X}}$ has not been reduced with further data $\langle \boldsymbol{y}_{\neq i_0} \rangle$. This simple updating approach establishes a m-dimensional inverse problem that is considered isolated from the remainder of the considered system, i.e. if other realizations $\boldsymbol{x}_i$ with $i \neq i_0$ are of inferential interest, analogous yet separate inverse problems have to be solved. A DAG-based visualization of the described situation is provided in Fig. 5.2. The flow of information towards $\boldsymbol{x}_{i_0}$ moves along the conditional relationships.

5.3.2 Staged estimation

In the second approach the additional data that were disregarded above can be processed. Initially the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ are inferred by probabilistic inversion of the data $\langle \boldsymbol{y}_{\neq i_0} \rangle$. The posterior $\pi(\boldsymbol{\theta}_{\boldsymbol{X}}|\langle \boldsymbol{y}_{\neq i_0} \rangle)$

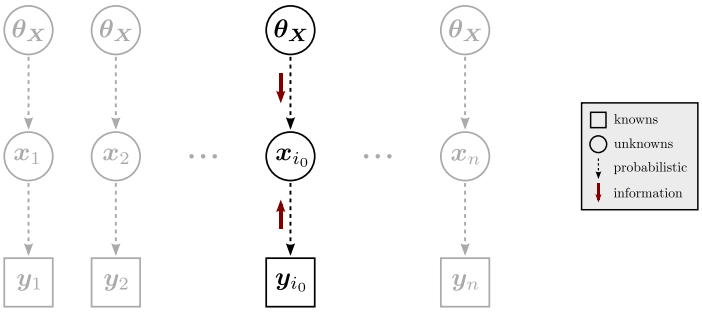


Figure 5.2: Simple updating.

obtained in this first step can be translated into the distribution

$$\pi(\boldsymbol{x}_{i_0}|\langle \boldsymbol{y}_{\neq i_0}\rangle) = \int f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_{i_0}|\boldsymbol{\theta}_{\boldsymbol{X}})\,\pi(\boldsymbol{\theta}_{\boldsymbol{X}}|\langle \boldsymbol{y}_{\neq i_0}\rangle)\,\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}. \tag{5.4}$$

It represents the uncertainty of $\boldsymbol{x}_{i_0}$ following the analysis of $\langle \boldsymbol{y}_{\neq i_0}\rangle$ but prior to analyzing $\boldsymbol{y}_{i_0}$. In a subsequent parameter estimation step $\pi(\boldsymbol{x}_{i_0}|\langle \boldsymbol{y}_{\neq i_0}\rangle)$ can be interpreted as a prior. The result of conditioning on $\boldsymbol{y}_{i_0}$ is a posterior distribution $\pi(\boldsymbol{x}_{i_0}|\langle \boldsymbol{y}_{\neq i_0}\rangle, \boldsymbol{y}_{i_0})$. In a sequential way the estimation of $\boldsymbol{x}_{i_0}$ has been based on the total number of data $\langle \boldsymbol{y}_i\rangle$. In the first stage the posterior of $\boldsymbol{\theta}_{\boldsymbol{X}}$ can be equivalently computed as the solution to a l-dimensional inference problem with a marginalized likelihood or as the marginal of the $(l+m\cdot(n-1))$-dimensional multilevel posterior $\pi(\langle \boldsymbol{x}_{\neq i_0}\rangle, \boldsymbol{\theta}_{\boldsymbol{X}}|\langle \boldsymbol{y}_{\neq i_0}\rangle)$. The second stage involves m-dimensional Bayesian updating of $\boldsymbol{x}_{i_0}$. The prior in Eq. (5.4) that is used in the second step contains a lower degree of uncertainty with respect to $\boldsymbol{x}_{i_0}$ than the one in Eq. (5.3). In this sense it is a "better" prior. The staged approach is visualized in Fig. 5.3 where initial probabilistic inversion is shown on the left, i.e. information accumulates at $\boldsymbol{\theta}_{\boldsymbol{X}}$. The subsequent updating step is shown on the right, i.e. information about $\boldsymbol{x}_{i_0}$ is extracted.

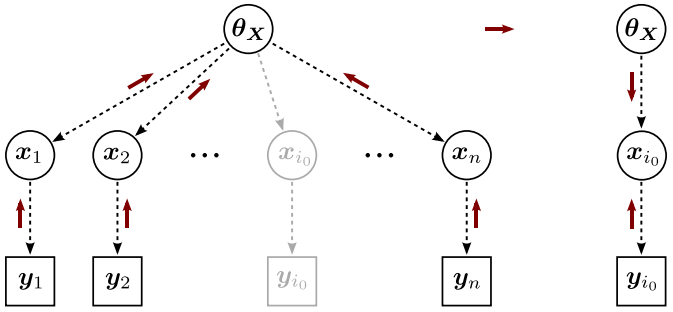


Figure 5.3: Staged estimation.

5.3.3 Multilevel inversion

Multilevel analysis of $\boldsymbol{x}_{i_0}$ is accomplished by constructing the joint posterior $\pi(\langle \boldsymbol{x}_i\rangle, \boldsymbol{\theta}_{\boldsymbol{X}}|\langle \boldsymbol{y}_i\rangle)$ in Eq. (5.2) and subsequently integrating out nuisance. Since inferential attention is not focused on the parameters $\langle \boldsymbol{x}_{\neq i_0}\rangle$ and hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$, those are marginalized out. The corresponding marginal of $\boldsymbol{x}_{i_0}$ is

$$\pi(\boldsymbol{x}_{i_0}|\langle \boldsymbol{y}_i\rangle) = \int \cdots \int \pi(\langle \boldsymbol{x}_i\rangle, \boldsymbol{\theta}_{\boldsymbol{X}}|\langle \boldsymbol{y}_i\rangle)\,\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}\,\mathrm{d}\langle \boldsymbol{x}_{\neq i_0}\rangle, \tag{5.5}$$

where the simplifying notation $\mathrm{d}\langle \boldsymbol{x}_{\neq i_0}\rangle = \mathrm{d}\boldsymbol{x}_1 \ldots \mathrm{d}\boldsymbol{x}_{i_0-1}\mathrm{d}\boldsymbol{x}_{i_0+1}\ldots\mathrm{d}\boldsymbol{x}_n$ is used. For estimating $\boldsymbol{x}_{i_0}$ the available information has been processed as a whole. Most notably the data $\langle \boldsymbol{y}_{\neq i_0}\rangle$ have been utilized for reducing the posterior uncertainty of $\boldsymbol{x}_{i_0}$. In practice, if the $(l+m\cdot n)$-dimensional posterior $\pi(\langle \boldsymbol{x}_i\rangle, \boldsymbol{\theta}_{\boldsymbol{X}}|\langle \boldsymbol{y}_i\rangle)$ is computed via an appropriate sampler, the marginal $\pi(\boldsymbol{x}_{i_0}|\langle \boldsymbol{y}_i\rangle)$ can be easily extracted by considering the corresponding $\boldsymbol{x}_{i_0}$-components only. The integral in Eq. (5.5) does not have to be computed explicitly. This way also other posterior marginals are obtained as a side product. Notwithstanding that the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ and realizations $\boldsymbol{x}_i$ other than $\boldsymbol{x}_{i_0}$ are not of immediate interest, they are incidentally inferred. In Fig. 5.4 a DAG-based illustration of the flow of information that governs the inference of $\boldsymbol{x}_{i_0}$ is shown. We remark that staged estimation and multilevel inversion formally resemble the Bayesian variants of filtering and smoothing [20], respectively, i.e. concepts from data assimilation in dynamical systems.

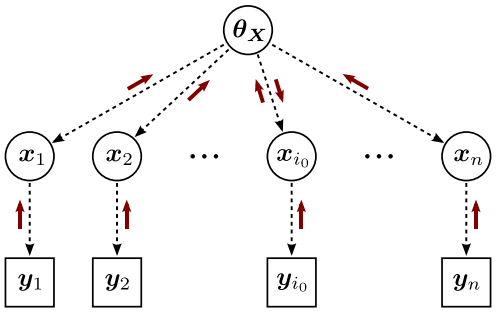


Figure 5.4: Multilevel inversion.

5.4 Hamiltonian Monte Carlo

In this section we shortly introduce the principles of *Hamiltonian Monte Carlo* or *hybrid Monte Carlo* (HMC) that was originally introduced in [16]. More comprehensive introductions can be found in [21–23]. HMC sampling allows one to obtain a widely decorrelated posterior sample by suppressing the diffusive random walk behavior of standard MCMC techniques. This is realized by embedding the parameter space in a higher-dimensional phase space and sampling from an appropriately defined auxiliary distribution that allows to extract the desired posterior as a marginal. The formulation draws on the Hamiltonian formalism of classical mechanics. More specifically it is inspired by a fictional classical particle moving in a potential well that is proportional to the negative log-density of the posterior. Candidate states are proposed following a dynamical simulation in the augmented state space. In doing so the search in the parameter space is guided by first-order derivative information from the posterior density. It allows for nonlocal MCMC moves that span whole regions of the parameter space that carry significant posterior probability mass.

HMC was originally developed for computational approaches to theoretical particle physics [16], where it accelerates the stochastic simulation of high-dimensional integrals [24]. Afterwards the potential for statistical applications was recognized [25]. Currently HMC has attracted greater attention in statistically and mathematically oriented scientific communities. Numerous extensions and generalizations have been proposed in the recent literature [26–29]. Notably there is the powerful yet costly Riemannian manifold HMC [30, 31]. HMC and its enhanced variants are applied in an increasing number of studies [32–34].

However, HMC-like algorithms are still widely underacknowledged for engineering problems down to the present day. Applications in structural dynamics and finite element modeling form exceptions [35, 36]. Likewise this holds for hierarchical statistical models. Although the software package Stan [37] offers an adaptive HMC-variant [38] for classical hierarchical models, i.e. without physical forward modeling, we are only aware of a very few studies wherein HMC is investigated [39, 40]. A semi-separable Hamiltonian for Riemannian manifold HMC sampling is proposed in [39]. Re-parametrizations of hierarchical models [41] in the context of HMC sampling are discussed in [40].

5.4.1 The MH algorithm

The principle of MCMC sampling is the construction of an ergodic Markov chain over the prior support that has the posterior as its stationary distribution. A prototypical class of MCMC techniques is based on the Metropolis-Hastings (MH) algorithm [42, 43]. More general introductions can be found in [14, 15].

Let $\pi_0(\boldsymbol{q})$ be the prior and $\pi_1(\boldsymbol{q}) \propto \mathcal{L}(\boldsymbol{q})\,\pi_0(\boldsymbol{q})$ the posterior of the unknown quantities $\boldsymbol{q} = (q_1, \ldots, q_d) \in \mathbb{R}^d$. The MH algorithm is started at an initial state $\boldsymbol{q}^{(0)}$ from the prior domain. Then it realizes a Markov chain with a long-run distribution $\pi_1(\boldsymbol{q})$ by repeatedly proceeding as follows. For a state $\boldsymbol{q}^{(t)}$ of the Markov chain in iteration t, a candidate state $\boldsymbol{q}^{(\star)} \sim P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})$ is sampled from an instrumental jumping distribution $P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})$. This proposal is then accepted as the new state with probability

$$\alpha\left(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)}\right) = \min\left\{1, \frac{\pi_1(\boldsymbol{q}^{(\star)})\,P(\boldsymbol{q}^{(t)}|\boldsymbol{q}^{(\star)})}{\pi_1(\boldsymbol{q}^{(t)})\,P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})}\right\}. \tag{5.6}$$

In this case the new state in iteration $t+1$ is $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(\star)}$. In case of rejection the chain remains in its state $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}$. The Markov chain transition kernel defined this way is easily seen to satisfy detailed balance with respect to the posterior. This is a sufficient condition for leaving the posterior invariant. An appealing feature of the MH correction Eq. (5.6) is that it only calls for evaluations of the unscaled posterior density. Moreover it gives ample scope for the design of efficient proposal distributions P.

A common MCMC updating scheme is the random walk Metropolis (RWM) sampler. It is based on local proposals that are sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{q}})$ with mean $\boldsymbol{q}^{(t)}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{q}}$. This leads to the classical diffusive random walk behavior. Note that for this symmetric proposals the acceptance probability in Eq. (5.6) reduces to $\alpha = \min\{1, \pi_1(\boldsymbol{q}^{(\star)})/\pi_1(\boldsymbol{q}^{(t)})\}$. Optimal scalings of the RWM in high-dimension are investigated in [44, 45].

5.4.2 Effective sample size

Due to the Markovian updates MCMC samples are generally autocorrelated. The autocorrelation governs the quality of the MCMC sample with respect to posterior expectations $\mu_g = \mathbb{E}[g(\boldsymbol{q})] = \int g(\boldsymbol{q})\pi_1(\boldsymbol{q})\,\mathrm{d}\boldsymbol{q}$ of a function of interest $g\colon \mathbb{R}^d \to \mathbb{R}$ [46, 47]. Under certain conditions one can show that the Markov chain $\boldsymbol{q}^{(0)}, \boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}, \ldots$ satisfies a central limit theorem [48, 49]. For a large number of iterations $N \to \infty$ the sample means $\bar{g} = N^{-1}\sum_{t=1}^{N} g_t$ with $g_t = g(\boldsymbol{q}^{(t)})$ approach a Gaussian distribution $\mathcal{N}(\bar{g}|\mu_{\bar{g}}, \sigma_{\bar{g}}^2)$ with mean $\mu_{\bar{g}} = \mu_g$ and asymptotic variance

$$\sigma_{\bar{g}}^2 = \frac{\sigma_g^2}{N}\left(1 + 2\sum_{s=1}^{\infty}\rho_s\right) = \frac{\sigma_g^2}{N}\tau_{\mathrm{int}} = \frac{\sigma_g^2}{N_{\mathrm{eff}}}. \tag{5.7}$$

Here the variance $\sigma_g^2 = \mathrm{Var}[g_t]$ and the lag-s autocorrelation $\rho_s = \mathrm{Cov}[g_t, g_{t+s}]/\sigma_g^2$ are statistical moments with respect to the stationary distribution. The *integrated autocorrelation time* in Eq. (5.7) is defined as $\tau_{\mathrm{int}} = 1 + 2\sum_{s=1}^{\infty}\rho_s$. Based on this one can define an *effective sample size* $N_{\mathrm{eff}} = N/\tau_{\mathrm{int}}$ which quantifies an equivalent number of independent draws from the posterior featuring the same standard error $\sigma_{\bar{g}} = \sigma_g/\sqrt{N_{\mathrm{eff}}}$ as the autocorrelated MCMC sample of size N. In this sense τ_{int} and N_{eff} are measures of the imprecision and effectiveness of simulating μ_g as $\bar{g}$, respectively. Note that with the projection $g\colon \boldsymbol{q} \mapsto q_j$ for $j \in \{1, \ldots, d\}$ these considerations straightforwardly apply to each posterior marginal.

5.4.3 Systems from classical physics

By analogy with two systems from classical physics, i.e. Newtonian and statistical mechanics [50, 51], the basic machinery of HMC is now outlined. We consider a hypothetical classical system with *canonical coordinates* $(\boldsymbol{q}, \boldsymbol{p})$, i.e. the *positions* $\boldsymbol{q} \in \mathbb{R}^d$ and *conjugate momenta* $\boldsymbol{p} \in \mathbb{R}^d$. Statistical QoI $\boldsymbol{q}$ are identified with positions of the system and momentum variables $\boldsymbol{p}$ are additionally introduced. The *Hamiltonian* of the system is given as

$$H(\boldsymbol{q}, \boldsymbol{p}) = V(\boldsymbol{q}) + T(\boldsymbol{p}), \tag{5.8}$$

where $V(\boldsymbol{q})$ and $T(\boldsymbol{p})$ are the *potential* and *kinetic energy*, respectively. The potential energy in Eq. (5.8) is defined as

$$V(\boldsymbol{q}) = -\log\left(\mathcal{L}(\boldsymbol{q})\,\pi_0(\boldsymbol{q})\right), \tag{5.9}$$

where $\pi_0(\boldsymbol{q})$ and $\mathcal{L}(\boldsymbol{q})$ are the prior density and the likelihood function, respectively. The kinetic energy term in Eq. (5.8) is defined as

$$T(\boldsymbol{p}) = \frac{\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p}}{2}, \tag{5.10}$$

where $\boldsymbol{M}$ is some symmetric and positive-definite *mass matrix*. Often it is a multiple $\boldsymbol{M} = m\boldsymbol{I}_d$ of the identity matrix $\boldsymbol{I}_d$ or of the general diagonal form $\boldsymbol{M} = \mathrm{diag}(m_1, \ldots, m_d)$.

As a first analogy to classical physics one considers *Hamiltonian dynamics* [50]. The evolution of the system in fictitious time τ is then governed by Hamilton's equations of motion (EoM)

$$\frac{\mathrm{d}\boldsymbol{q}}{\mathrm{d}\tau} = \frac{\partial H}{\partial \boldsymbol{p}}, \quad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}\tau} = -\frac{\partial H}{\partial \boldsymbol{q}}. \tag{5.11}$$

The governing differential equations in Eq. (5.11) determine the system trajectory over time. This dynamics satisfies time reversibility (invariance of the dynamics under the transformation $(\tau, \boldsymbol{p}) \mapsto (-\tau, -\boldsymbol{p})$), energy conservation ($\mathrm{d}H/\mathrm{d}\tau = 0$) and preservation of the phase space volume (Liouville's theorem).

As a second analogy to classical physics one considers the distribution of the *canonical ensemble* from statistical mechanics [51]. The fictitious temperature and the Boltzmann constant are set to one. Then the frequency distribution of the positions and the momenta is the *Boltzmann distribution*

$$\Pi_1(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{Z}\exp(-H(\boldsymbol{q}, \boldsymbol{p})) = \frac{\mathcal{L}(\boldsymbol{q})\,\pi_0(\boldsymbol{q})}{Z}\,\exp\left(-\frac{\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p}}{2}\right). \tag{5.12}$$

Here the normalizing constant Z is the *canonical partition function*. It does not have to be explicitly known for HMC sampling. By construction, i.e. due to the definitions in Eqs. (5.8) to (5.10), the joint distribution Eq. (5.12) has the form $\Pi_1(\boldsymbol{q},\boldsymbol{p}) = \pi_1(\boldsymbol{q})\,\pi_1(\boldsymbol{p})$. It features the posterior $\pi_1(\boldsymbol{q}) \propto \exp(-V(\boldsymbol{q})) = \mathcal{L}(\boldsymbol{q})\,\pi_0(\boldsymbol{q})$ as its $\boldsymbol{q}$-marginal. Moreover the $\boldsymbol{p}$-marginal of Eq. (5.12) is a multivariate Gaussian $\pi_1(\boldsymbol{p}) \propto \exp(-T(\boldsymbol{p})) = \exp(\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p}/2)$ with covariance matrix $\boldsymbol{M}$. The partition function Z absorbs the normalization factors of $\pi_1(\boldsymbol{q})$ and $\pi_1(\boldsymbol{p})$.

5.4.4 The HMC algorithm

The core idea of HMC is to realize an ergodic Markov chain over the $2d$-dimensional configuration space of the classical system that was defined above. This chain is constructed in such a way that it features Eq. (5.12) as its stationary distribution, with the posterior as a marginal. After the HMC algorithm is initialized at a certain $\boldsymbol{q}^{(0)}$ one iteratively applies the following Markovian transition. Given the current positions $\boldsymbol{q}^{(t)}$ of the Markov chain at iteration t, the corresponding momenta are directly sampled from the distribution $\pi_1(\boldsymbol{p}^{(t)}) = \mathcal{N}(\boldsymbol{p}^{(t)}|\boldsymbol{0},\boldsymbol{M})$. The system configuration $(\boldsymbol{q}^{(t)},\boldsymbol{p}^{(t)})$ is then evolved over some arbitrary time interval, after which it has reached a new configuration $(\boldsymbol{q}^{(\star)},\boldsymbol{p}^{(\star)})$. Following this, the updated position $\boldsymbol{q}^{(\star)}$ is the new state $\boldsymbol{q}^{(t+1)}$ of the Markov chain in iteration $t+1$, i.e. acceptance by default. Since we are only interested in positions, the auxiliary momentum $\boldsymbol{p}^{(\star)}$ is discarded. Time reversibility, the conservation of energy and the preservation phase space volume are important dynamical properties of the EoM in Eq. (5.11). Based on the latter two properties one can show that the transition defined above leaves the Boltzmann distribution Eq. (5.12) invariant [23].

The abovementioned ideal HMC updating scheme avoids the symmetric and strongly localized proposals of RWM-type algorithms. Properly tuned the dynamical transitions may cover wide regions of the position space that accumulate significant posterior mass. It can be extremely efficient for sampling high-dimensional and strongly correlated posterior distributions. This makes HMC a promising candidate sampler for hierarchical models that are higher-dimensional and correlated per definition.

In practice idealized HMC updating based on exactly solving the EoM cannot be accomplished. Instead one has to resort to numerical simulations of Hamiltonian dynamics based on suitable integrators [52, 53]. A standard choice is the *leapfrog* time-stepping scheme [54], but note that also other symplectic integrators could be used [55]. The system is evolved from its configuration at time τ_1 into the one at $\tau_2 > \tau_1$ by an iterative computation of the position and momentum variables

$$\boldsymbol{q}(\tau+\Delta\tau) = \boldsymbol{q}(\tau) + \Delta\tau \boldsymbol{M}^{-1}\boldsymbol{p}(\tau+\frac{1}{2}\Delta\tau), \tag{5.13a}$$

$$\boldsymbol{p}(\tau+\frac{3}{2}\Delta\tau) = \boldsymbol{p}(\tau+\frac{1}{2}\Delta\tau) - \Delta\tau\frac{\partial V}{\partial \boldsymbol{q}}(\boldsymbol{q}(\tau+\Delta\tau)). \tag{5.13b}$$

Starting from $(\boldsymbol{q}(\tau_1),\boldsymbol{p}(\tau_1)) \equiv (\boldsymbol{q}^{(t)},\boldsymbol{p}^{(t)})$ the momentum $\boldsymbol{p}(\tau_1+\frac{1}{2}\Delta\tau) = \boldsymbol{p}(\tau_1) - \frac{1}{2}\Delta\tau\frac{\partial V}{\partial \boldsymbol{q}}(\boldsymbol{q}(\tau_1))$ is computed in a half step. Thereafter alternating full steps are done for the positions and momenta according to Eq. (5.13). Finally a half step is taken from $\boldsymbol{p}(\tau_2-\frac{1}{2}\Delta\tau)$ to $\boldsymbol{p}(\tau_2) = \boldsymbol{p}(\tau_2-\frac{1}{2}\Delta\tau) - \frac{1}{2}\Delta\tau\frac{\partial V}{\partial \boldsymbol{q}}(\boldsymbol{q}(\tau_2))$. At the end the system has evolved into $(\boldsymbol{q}(\tau_2),\boldsymbol{p}(\tau_2)) \equiv (\boldsymbol{q}^{(\star)},\boldsymbol{p}^{(\star)})$. This way the computation over the time interval $\tau_2-\tau_1 = L\Delta\tau$ has been discretized into a discrete number of steps L with a finite stepsize $\Delta\tau$. A visualization of this time integration scheme is provided in Fig. 5.5. The full and half steps of the time evolution of the system are shown.

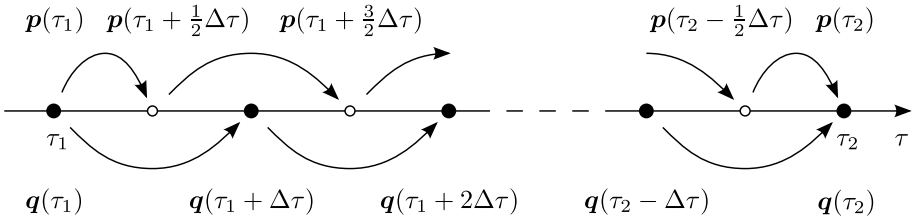


Figure 5.5: Leapfrog time stepping.

An appealing property of the leapfrog is that it approximates Hamiltonian dynamics in a way that exactly maintains time reversibility and volume preservation. The total energy is approximately conserved with an error that asymptotically is of the order $\mathcal{O}(\Delta\tau^2)$. This introduces a characteristic scale of the stepsize that is related to stable trajectories. In order to compensate for the introduced approximation of Hamiltonian dynamics, candidate configurations $(\boldsymbol{q}^{(\star)},\boldsymbol{p}^{(\star)})$ are accepted with probability

$$\alpha\left(\boldsymbol{p}^{(\star)},\boldsymbol{q}^{(\star)}|\boldsymbol{p}^{(t)},\boldsymbol{q}^{(t)}\right) = \min\left\{1, \exp\left(H\left(\boldsymbol{p}^{(t)},\boldsymbol{q}^{(t)}\right) - H\left(\boldsymbol{p}^{(\star)},\boldsymbol{q}^{(\star)}\right)\right)\right\}. \tag{5.14}$$

This is plain vanilla Metropolis correction in the $2d$-dimensional phase space for a symmetric proposal distribution [23]. Volume preservation and time reversibility of the leapfrog integration are the properties that ensure the symmetry in the proposals. Strictly speaking one would have to negate the momenta at the end of the trajectory, however, those are disregarded anyhow. Due to Eq. (5.14) the acceptance rate depends on the degree as to which energy conservation is violated. The combined transition satisfies detailed balance with respect to the Boltzmann distribution Eq. (5.12). Thus the Markov chain $\boldsymbol{q}^{(0)}, \boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}, \dots$ exhibits the stationary distribution $\pi_1(\boldsymbol{q})$. Notice that, similar to the MH algorithm, HMC sampling requires only evaluations of the unnormalized posterior density.

Additionally the time integration in Eq. (5.13) requires the computation of first-order partial derivatives of the unscaled posterior log-density in Eq. (5.9). In turn this requires the differentiation of the forward model with respect to its inputs. For simple models this can be done analytically [56]. Otherwise one has to rely on the adjoint method [57], automatic differentiation [35] or the straightforward use of finite differences [36]. It is interesting to note that derivatives do not have to be calculated exactly. Although this may decrease the acceptance rate, such approximations are managed in the correction step Eq. (5.14). An interesting idea would be to employ such approximations of the forward model for which derivatives can be analytically obtained, e.g. polynomial chaos metamodels [58].

Other practical issues relate to the handling of parameter constraints [59, 60] and the optimal tuning of the algorithm [61, 62]. Free algorithmic parameters of the HMC are the number of leapfrog steps L and the timestep $\Delta\tau$. Together they determine the total trajectory length $L\Delta\tau$. Furthermore the mass matrix $\boldsymbol{M}$ has to be set. The latter is often chosen to be a diagonal matrix $\boldsymbol{M} = \mathrm{diag}(m_1, \dots, m_d)$. Individually setting the entries m_i for $i = 1, \dots, d$ then allows to account for different posterior scales of q_i, e.g. with $m_i = 1/\mathrm{Var}[q_i]$ where $\mathrm{Var}[q_i]$ is the marginal posterior variance. A more in-depth discussion of related issues is found in [22].

5.5 Numerical experiments

In order to demonstrate the optimal inference of individual parameters, we devise a simple example within the domain of structural engineering for which we conduct a simulated computer experiment. It should be understood as a benchmark of optimal combination of information in data analysis of engineering systems. An experimental situation is investigated where data are collected for an ensemble of identically manufactured beams. The acquired specimen-specific data are informative to variable degree. Therefore the goal is to optimally exploit the available information in the individual assessment of ensemble members that are poorly supported by experimental evidence.

The system under consideration is a set of simply supported beams $i = 1, \dots, n$ with well-known lengths L_i, widths b_i and heights h_i. Beams are composed out of a material which is subject to aleatory uncertainty in its material properties, say the Young's modulus E_i. For each individual beam i the elastic modulus E_i is assumed to be constant along its main axis. Across the sample of beams Young's moduli E_i are distributed according to a lognormal distribution $\mathcal{LN}(E_i|\mu_E, \sigma_E^2)$ with mean value $\mu_E = \mathbb{E}[E_i]$ and variance $\sigma_E^2 = \mathrm{Var}[E_i]$. At positions s_j with $j = 1, \dots, n_i$ and $0 \leq s_j \leq L_i/2$ the deflections $v_i(s_j)$ of the beams under concentrated point loads F_i at midspan are calculated as

$$v_i(s_j) = \frac{F_i s_j}{48 E_i I_i} \left(3L_i^2 - 4s_j^2\right). \tag{5.15}$$

Here the moment of inertia is given as $I_i = b_i h_i^3/12$. A symmetric expression holds for positions s_j with $L_i/2 \leq s_j \leq L_i$. In Fig. 5.6 a sketch of a simply supported beam is drawn. In a series of experiments measured beam deflections can be used to estimate individual realizations E_i or the hyperparameters (μ_E, σ_E). Herein we consider the inference of the Young's modulus E_{i_0} of a beam $i_0 \in \{1, \dots, n\}$ for which it is assumed that experimental evidence is scarce. A simulated computer experiment is conducted as described below.

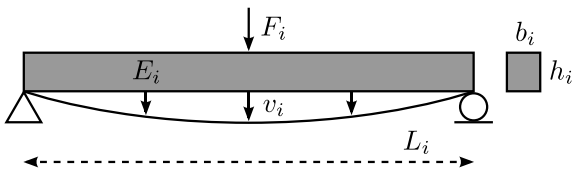


Figure 5.6: Simply supported beam.

We choose a set of $n = 100$ beams with well-known and constant dimensions $L_i = 1\,\mathrm{m}$ and $b_i = h_i = 10\,\mathrm{cm}$ that are subjected to loads $F_i = 30\,\mathrm{kN}$. The elastic moduli E_i are randomly sampled from a lognormal distribution with mean $\mu_E = 15\,\mathrm{GPa}$ and standard deviation $\sigma_E = 3\,\mathrm{GPa}$. We simulate a synthetic set of pseudo-data

$\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})$ for each beam. For $j = 1, 2, 3$ pseudo-observations $y_{ij} = v_i(s_j) + \epsilon_{ij}$ are generated for positions $\boldsymbol{s}_i = (s_1, s_2, s_3)$ with $s_1 = 25\,\text{cm}$, $s_2 = 50\,\text{cm}$ and $s_3 = 75\,\text{cm}$ by perturbing the corresponding model predictions Eq. (5.15). The residuals $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})$ are independently sampled from centered Gaussians $\mathcal{N}(\boldsymbol{\varepsilon}_i|\mathbf{0}, \boldsymbol{\Sigma}_i)$ with covariance matrices $\boldsymbol{\Sigma}_i = \text{diag}(\sigma_{i1}^2, \sigma_{i2}^2, \sigma_{i3}^2)$. We choose comparably small residual standard deviations $\sigma_{ij} = 0.01\,\text{cm}$ for $i \neq i_0$ and comparably large deviations $\sigma_{i_0 j} = 0.1\,\text{cm}$. This represents an experimental situation where the ensemble of beams is tested, however, the data are less informative about the specimen i_0.

We assume that the hyperparameter values (μ_E, σ_E) are not known, yet prior or expert knowledge is available. The inferential prior of the hyperparameters is set as the independent distribution $\pi(\mu_E, \sigma_E) = \pi(\mu_E)\,\pi(\sigma_E)$ with weakly informative but proper uniform distributions $\pi(\mu_E) = \mathcal{U}(\mu_E|0, 100)$ and $\pi(\sigma_E) \sim \mathcal{U}(\sigma_E|0, 30)$ (in GPa) as marginals. This represents the scenario that the available prior knowledge does not allow to assign more informative priors. Nonetheless the hyperparameter values are physically bounded by zero from below and they can be priorly bounded from above. Thus values that are outside of the specified interval can be excluded.

In the following we conduct inference of the Young's modulus E_{i_0} of a beam i_0. The hyperparameter values (μ_E, σ_E) and the randomly sampled elastic moduli $\langle E_i \rangle$ are treated as "unknowns". This includes the "true" parameter value $E_{i_0} = 17.01\,\text{GPa}$. The measurement locations $\langle \boldsymbol{s}_i \rangle$, applied loads $\langle F_i \rangle$ and physical beam dimensions $\langle L_i, b_i, h_i \rangle$ are the well-known experimental conditions. Further knowns comprise the parametric prior $\pi(\mu_E, \sigma_E)$, the levels of measurement uncertainty $\langle \boldsymbol{\Sigma}_i \rangle$ and the data $\langle \boldsymbol{y}_i \rangle$. The described experimental setup is studied next. Simple Bayesian updating, staged estimation and multilevel inversion are demonstrated. The flow of information is investigated and insight into the inferential mechanism is provided. Low-dimensional posteriors are generally computed by means of a RWM algorithm, while higher-dimensional posteriors are computed by means of HMC/RWM hybrid sampler. The latter is based on updating the parameters via Hamiltonian dynamics while updating the hyperparameters with a random walk. Finally the implementation of the samplers is described in detail. Furthermore, the efficiency of HMC/RWM sampling is contrasted with pure RWM sampling.

5.5.1 Simple updating

The compound prior Eq. (5.3) may not be available in analytical form, however, one can sample the prior by the method of composition [63]. To that end one draws K samples from the hyperparameter distribution $\pi(\mu_E, \sigma_E)$. Subsequently one draws a sample from the parameter distribution $\mathcal{LN}(E_i|\mu_E, \sigma_E^2)$ for each hyperparameter realization. As desired the sample of parameter values is distributed according to the mixture in Eq. (5.3). We draw $K = 10^5$ random samples of the mixture prior. In Fig. 5.7 a kernel density estimate of the mixture is shown. It is seen that the mixture is approximated well by kernel smoothing. In the following MCMC analysis the prior of E_{i_0} is therefore represented as the obtained kernel density estimate. We accomplish Bayesian updating by sampling the one-dimensional posterior $\pi(E_{i_0}|\boldsymbol{y}_{i_0})$ with a simple RWM sampler. In Fig. 5.8 the simulated posterior is shown. The posterior of E_{i_0} has the mean $\mu_{E_{i_0}} = 21.73\,\text{GPa}$ and the standard deviation $\sigma_{E_{i_0}} = 5.49\,\text{GPa}$. With a coefficient of variation $\text{CV} = 25\,\%$ this corresponds to a relatively high degree of posterior uncertainty. The maximum a posteriori (MAP) estimate of the modulus E_{i_0}, i.e. the posterior mode, is found to be $\hat{E}_{i_0}^{\text{MAP}} = 19.07\,\text{GPa}$. Compared to the true value $E_{i_0} = 17.01\,\text{GPa}$ the relative approximation error of the MAP estimate is $\epsilon = 12\,\%$.

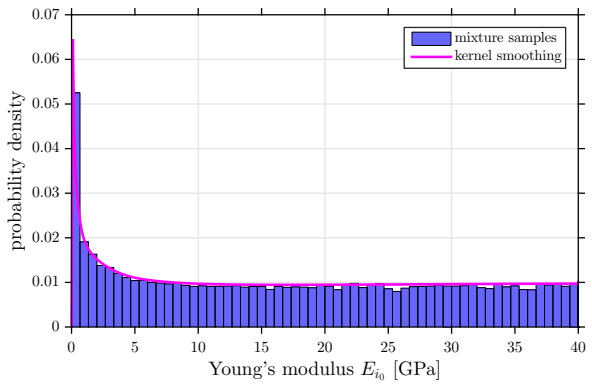


Figure 5.7: Simple updating: prior.

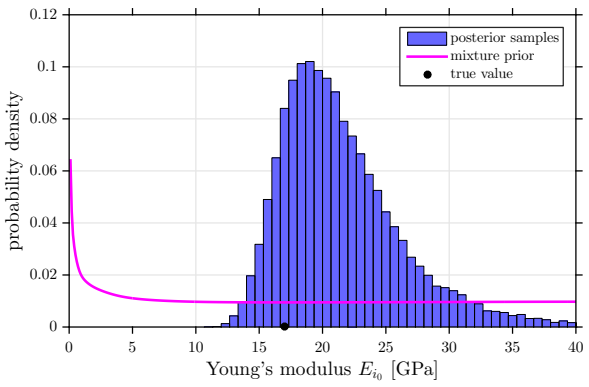


Figure 5.8: Simple updating: posterior.

5.5.2 Staged estimation

First we conduct probabilistic inversion to infer the hyperparameters (μ_E, σ_E) with data $\langle \boldsymbol{y}_{\neq i_0} \rangle$. The high-dimensional posterior $\pi(\langle E_{\neq i_0} \rangle, \mu_E, \sigma_E | \langle \boldsymbol{y}_{\neq i_0} \rangle)$ is therefore computed via HMC sampling. The implementation of the sampler is described later on. For the time being we remark that $N = 10^6$ posterior samples are drawn within a runtime of ca. $t = 1\,\mathrm{h}$. A thinned sample of the posterior $\pi(\mu_E, \sigma_E | \langle \boldsymbol{y}_{\neq i_0} \rangle)$ is used to draw $K = 10^5$ samples from Eq. (5.4) by the composition method. The resulting sample from $\pi(E_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle)$ is shown in Fig. 5.9 along with a lognormal fit to the sample. Since the lognormal fit reproduces the desired distribution adequately well, it is utilized as the prior $\pi(E_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle)$ in the subsequent updating step. The resulting univariate posterior $\pi(E_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle, \boldsymbol{y}_{i_0})$ is shown in Fig. 5.10. The final posterior of E_{i_0} has the mean $\mu_{E_{i_0}} = 17.44\,\mathrm{GPa}$ and the standard deviation $\sigma_{E_{i_0}} = 2.17\,\mathrm{GPa}$. With $\mathrm{CV} = 12\,\%$ the posterior features a lower degree of uncertainty as compared to simple updating. Likewise the relative error $\epsilon = 1\,\%$ of the MAP estimate $\hat{E}_{i_0}^{\mathrm{MAP}} = 16.84\,\mathrm{GPa}$ is much smaller.

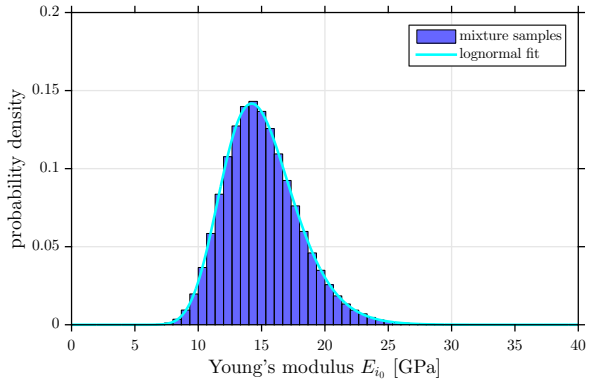


Figure 5.9: Staged estimation: prior.

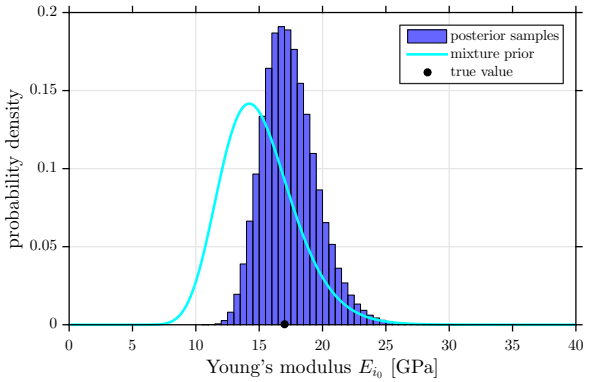


Figure 5.10: Staged estimation: posterior.

Summarized, the posterior $\pi(E_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle, \boldsymbol{y}_{i_0})$ shown in Fig. 5.10 is a better representation of the true value than $\pi(E_{i_0} | \boldsymbol{y}_{i_0})$ depicted in Fig. 5.8. The reason is that the prior $\pi(E_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle)$ plotted in Fig. 5.9 contains more information than $\pi(E_{i_0})$ shown in Fig. 5.7. This illustrates the flow of information from data $\langle \boldsymbol{y}_{\neq i_0} \rangle$ towards E_{i_0} that was already discussed. The information exchange takes place in an indirect way. While staged estimation requires the consecutive solution of two different problems, i.e. probabilistic inversion and subsequent updating, multilevel inversion is more elegant and satisfactory in the sense that it allows to consistently perform those two separate tasks at once. This is demonstrated next.

5.5.3 Multilevel inversion

Full multilevel inversion is finally performed by sampling the joint posterior $\pi(\langle E_i \rangle, \mu_E, \sigma_E | \langle \boldsymbol{y}_i \rangle)$. The employed MCMC sampler is described and benchmarked afterwards. Samples from the marginal Eq. (5.5) can easily be extracted by discarding samples from $\langle E_{\neq i_0} \rangle$ and (μ_E, σ_E). The marginal posterior $\pi(E_{i_0} | \langle \boldsymbol{y}_i \rangle)$ is plotted in Fig. 5.11. It has the mean $\mu_{E_{i_0}} = 17.44\,\mathrm{GPa}$ and the standard deviation $\sigma_{E_{i_0}} = 2.18\,\mathrm{GPa}$ which rounds up to $\mathrm{CV} = 13\,\%$. The relative error of the MAP estimate $\hat{E}_{i_0}^{\mathrm{MAP}} = 16.86\,\mathrm{GPa}$ is $\epsilon = 1\,\%$. As a summary the simulated posteriors $\pi(E_{i_0} | \boldsymbol{y}_{i_0})$, $\pi(E_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle, \boldsymbol{y}_{i_0})$ and $\pi(E_{i_0} | \langle \boldsymbol{y}_i \rangle)$ that are relevant to the identification of E_{i_0} are shown in Fig. 5.12. The corresponding means, modes and standard deviations (SD) are listed in Table 5.1.

Table 5.1: Summary of estimating E_{i_0}.

Approach	Mean [GPa]	Mode [GPa]	SD [GPa]
Simple updating	21.73	19.07	5.49
Staged estimation	17.44	16.84	2.17
Multilevel inversion	17.44	16.86	2.18

As expected, the posterior $\pi(E_{i_0} | \langle \boldsymbol{y}_i \rangle)$ shown in Fig. 5.11 is nearly identical to $\pi(E_{i_0} | \langle \boldsymbol{y}_{\neq i_0} \rangle, \boldsymbol{y}_{i_0})$ in Fig. 5.10. By comparison with the corresponding posterior $\pi(E_{i_0} | \boldsymbol{y}_{i_0})$ in Fig. 5.8, the additional amount of information relating to E_{i_0} that has been gained by utilizing the full set of observations $\langle \boldsymbol{y}_i \rangle$ becomes apparent from the

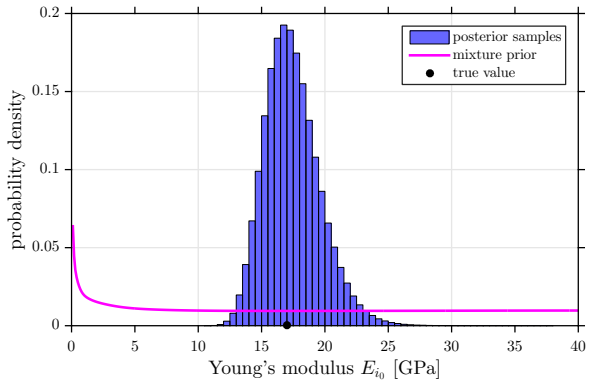


Figure 5.11: Multilevel inversion: posterior.

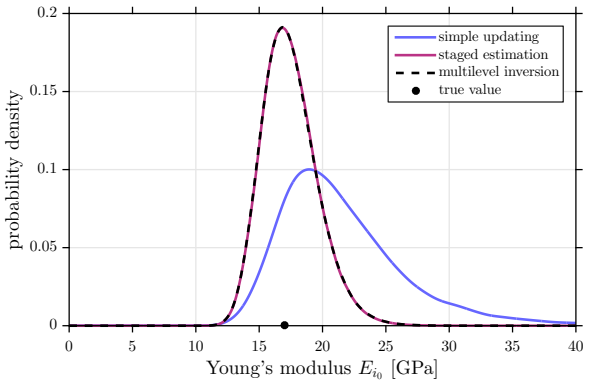


Figure 5.12: Summary of the posteriors.

associated shrinkage of the posterior uncertainty. A manifest advantage of multilevel inversion over staged estimation is that it incidentally provides information about all unknowns $(\langle E_i \rangle, \mu_E, \sigma_E)$. The posterior $\pi(\langle E_i \rangle, \mu_E, \sigma_E | \langle \boldsymbol{y}_i \rangle)$ accumulates information such that the whole data $\langle \boldsymbol{y}_i \rangle$ contributes to the collective learning process. This is advantageous in the case that not just a single realization E_{i_0} but a larger subset of $\langle E_i \rangle$ is of inferential interest. While staged estimation mainly served the purpose of illustrating the presence and manner of the abovementioned learning process, it is practical in experimental situations where data are not collected at the same time. In this case the sequential approach allows to analyze newly observed data without the need for resolving the full multilevel problem again.

5.5.4 Algorithmic efficiency

The efficacy and viability of HMC sampling for exploring multilevel posteriors is now investigated. In the last two sections the algorithm was used to compute the posteriors $\pi(\langle E_{\neq i_0} \rangle, \mu_E, \sigma_E | \langle \boldsymbol{y}_{\neq i_0} \rangle)$ and $\pi(\langle E_i \rangle, \mu_E, \sigma_E | \langle \boldsymbol{y}_i \rangle)$. The latter involves a 102-dimensional parameter space. In order to assess the performance gain by HMC sampling the algorithm is compared to a classical RWM algorithm. The RMW sampler is implemented in a blocked manner where the block of hyperparameters and the one of the parameters are updated separately. Gaussian proposal distributions with covariance matrices $\boldsymbol{\Sigma}_{(\mu_E, \sigma_E)} = \mathrm{diag}(0.3, 0.3)$ and $\boldsymbol{\Sigma}_{\langle E_i \rangle} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ are used. The latter is defined by setting $\sigma_{i_0} = 0.3$ and $\sigma_i = 0.002$ for $i \neq i_0$. This takes the different scales of the posterior marginals into account. The sampler is initialized at maximum likelihood estimates of individual parameters in separate deterministic inverse problems and two-stage estimates of the hyperparameters. This procedure initializes the algorithm close to the posterior modes. The MCMC execution time to draw $N = 10^7$ posterior samples amounts to ca. $t = 5\,\mathrm{h}$. Acceptance rates of ca. 22 % for the parameters and ca. 25 % for the hyperparameters are noticed.

The HMC/RWM sampler features the same blockwise updating structure. Hyperparameters are updated by the same random walk updates that were described above. The parameters are updated with HMC proposals. For the present problem all necessary partial derivatives are analytically obtained. A diagonal mass matrix $\boldsymbol{M} = \mathrm{diag}(m_1, \ldots, m_n)$ is used. The point masses are set to $m_i = 20$ for $i \neq i_0$ and to $m_{i_0} = 0.15$ for the QoI. This tuning accounts for the different marginal scales of the parameters. More details about relative parameter scalings can be found in [22]. For the dynamical simulation the parameters are set as follows. The number of leapfrog steps is set to $L = 8$. At the start of the trajectory the discretized stepsize $\Delta\tau$ in fictional time is randomized within the interval $[0.15, 0.16]$. It is kept fixed throughout each dynamical simulation, though. This avoids potentially occurring problems such as slow mixing or even of non-ergodicity due to (nearly) exact periodicity of the system trajectories [64]. The used parameter tuning leads to stable trajectories while it also ensures that long distances in the parameter space are traversed. Initialization is accomplished as before. A number of $N = 10^6$ MCMC iterations are executed within an execution runtime of ca. $t = 1\,\mathrm{h}$. The dynamical HMC parameter updates are accepted with a rate of ca. 100 %. In the hyperparameter block the acceptance rate is ca. 25 %.

The Markov chains produced by the RWM and the HMC sampler are compared with each other. In Figs. 5.13 and 5.14 the converged chains for E_{i_0} are shown for 5000 MCMC iterations. These trace plots show how the chains sample the corresponding posterior marginals around their means. Obviously the mixing properties of the HMC chain are better than the ones of the RWM chain. The sample autocorrelation function (ACF) gives greater insight into the characteristics of the samplers. In Figs. 5.15 and 5.16 the ACFs of the two chains are shown for the QoI E_{i_0}. It can be seen that for RWM sampling the ACF drops to zero within ca. 1000 MCMC

iterations. In contrast the ACF of the HMC chain dies down within ca. five iterations.

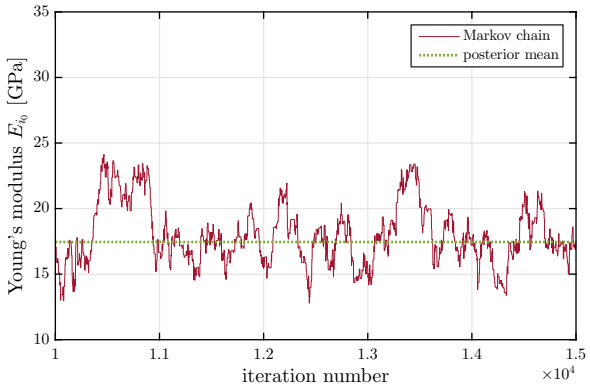


Figure 5.13: RWM: Trace plot of E_{i_0}.

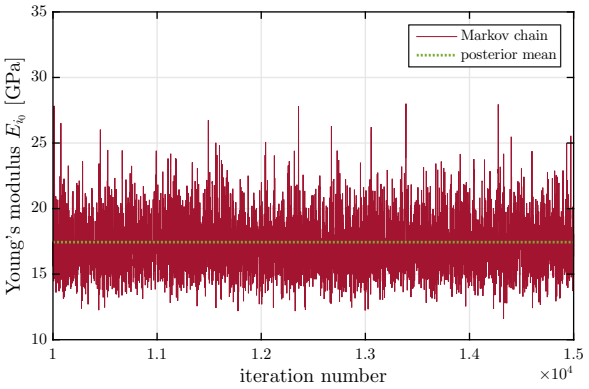


Figure 5.14: HMC: Trace plot of E_{i_0}.

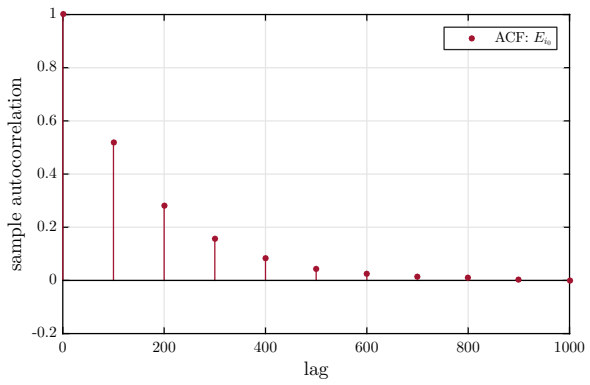


Figure 5.15: RWM: ACF of E_{i_0}.

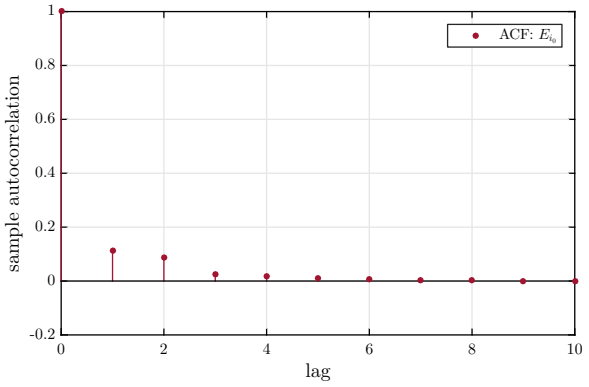


Figure 5.16: HMC: ACF of E_{i_0}.

With the ACF one can approximately assess the effective sample size N_{eff} in Eq. (5.7). A variety of methods for assessing the estimation variance $\sigma^2_{\bar{E}_{i_0}} = N^{-1}\sigma^2_{E_{i_0}}\tau_{\text{int}}$ and the effective sample size $N_{\text{eff}} = N/\tau_{\text{int}}$ have been proposed [65, 66]. The simplest method rests upon the substitution of $\sigma^2_{E_{i_0}}$ and τ_{int} by estimators [67, 68]. We use a sample estimate $\hat{\sigma}^2_{E_{i_0}}$ and a truncated series approximation $\hat{\tau}_{\text{int}} = 1 + 2\sum_{s=1}^{r}\hat{\rho}_s$. The sum is cut off at the smallest lag r for which the absolute value of the empirical autocorrelation $\hat{\rho}_r$ drops below a certain threshold $|\hat{\rho}_r| < 0.01$. For the RWM chain with a total number of iterations $N = 10^7$ that were done in $t = 5\,\text{h}$ one obtains an effective sample size $N_{\text{eff}} = 3 \cdot 10^4$ and a Monte Carlo standard error (MCSE) $\sigma_{\bar{E}_{i_0}} = 1.5 \cdot 10^{-2}\,\text{GPa}$ for estimating E_{i_0}. The effective sample size and the MCSE for the HMC chain with $N = 10^6$ and $t = 1\,\text{h}$ equal to $N_{\text{eff}} = 7 \cdot 10^5$ and $\sigma_{\bar{E}_{i_0}} = 2.5 \cdot 10^{-3}\,\text{GPa}$, respectively. A summary of those rounded values is given in Table 5.2. This amounts to a considerable speedup of about two orders of magnitude.

Table 5.2: Summary of sampling E_{i_0}.

Algorithm	N	N_{eff}	MCSE	t	N_{eff}/t
RWM	10^7	$3 \cdot 10^4$	$1.5 \cdot 10^{-2}\,\text{GPa}$	$5\,\text{h}$	$6 \cdot 10^3\,\text{h}^{-1}$
HMC/RWM	10^6	$7 \cdot 10^5$	$2.5 \cdot 10^{-3}\,\text{GPa}$	$1\,\text{h}$	$7 \cdot 10^5\,\text{h}^{-1}$

We observe that parameters different from E_{i_0} show similar mixing properties. In Figs. 5.17 and 5.18 the ACFs of the HMC series are exemplarily shown for two other parameters E_{i_1} and E_{i_2}. For the former the ACF vanishes almost instantaneously, which implies perfectly decorrelated updates. For the latter the ACF alternates between positive and negative values before it eventually vanishes within ca. five MCMC iterations. While for a reversible Markov chain the sum $\rho_{2t} + \rho_{2t+1}$ of autocorrelations at even-lag $2t$ and the adjacent odd-lag $2t+1$ must be strictly positive [46], negative autocorrelations cannot be ruled out principally. However, they typically do not occur for RWM updating schemes. For the nonlocal updates of HMC-like samplers, negative odd-lag autocorrelations may indicate that the trajectory is too long. If the dynamical simulation starts above/below the

posterior mean, then the trajectory tends to end below/above the mean. Altogether these observations suggest that the overall performance of the HMC algorithm can be further improved, e.g. through a fine tuning of the mass matrix.

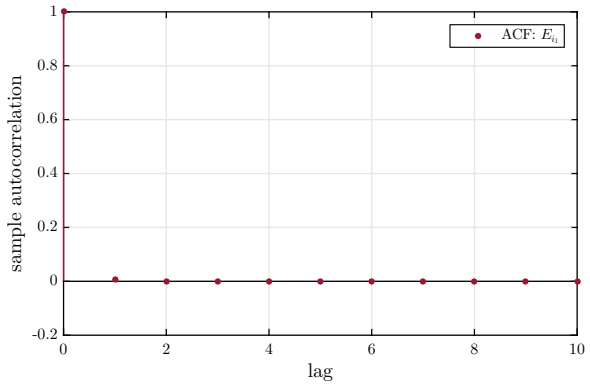


Figure 5.17: HMC: ACF of E_{i_1}.

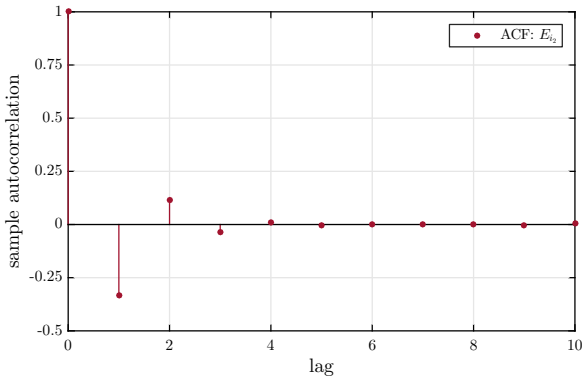


Figure 5.18: HMC: ACF of E_{i_2}.

5.6 Concluding remarks

Simple updating, staged estimation and multilevel inversion were proposed as Bayesian approaches to uncertainty reduction and information gathering in hierarchically defined inversed problems. Multilevel inversion was shown to outclass simple updating and staged estimation in terms of consistency and effectiveness. Specifically it facilitates the coherent inference of the problem unknowns where available prior information and experimental data are optimally combined. In the case that intermediate model variables are of inferential interest, this allows for pooling statistical strength. The potential of borrowing strength was demonstrated for the optimal estimation of material properties. The mutual exchange of information between specimens from a statistical ensemble was investigated. Ensemble members that are subject to a high degree of uncertainty can be advantageously assessed by exploiting information from members that are more strongly informed by the data. In our example application this allowed to mitigate the influence of high measurement uncertainty.

Moreover posterior exploration in high-dimensional parameter spaces was addressed. HMC was proven to be a practical and highly efficient sampler for hierarchical problems. For the system under consideration it outperforms RWM by two orders of magnitude as measured by the number of effective posterior samples that can be simulated within a given execution time. Put another way, for achieving a certain number of effective draws it reduces the execution time by two orders of magnitude. That way the HMC algorithm enables uncertainty quantification in more complex problems where the employment of classical MCMC techniques would be unfeasible. The high computational cost associated with traditional algorithms may easily exceed the available budget for problems that involve more sophisticated representations of uncertainty or more resource intensive forward models.

With increasing dimensionality of the parameter space, e.g. due to a refined uncertainty model, HMC promises to yield relative speedups that are even higher than the one observed in our benchmark application. For more advanced forward models, efficient and sufficiently accurate means to evaluate their derivatives have to be devised. A promising idea is the use of surrogate models such as polynomial chaos expansions. Research in this direction is in progress.

References

[1] R. Hadidi and N. Gucunski. “Probabilistic Approach to the Solution of Inverse Problems in Civil Engineering”. In: *Journal of Computing in Civil Engineering* 22.6 (2008), pp. 338–347. DOI: 10.1061/(ASCE)0887-3801(2008)22:6(338).

[2] J. L. Beck. “Bayesian system identification based on probability logic”. In: *Structural Control and Health Monitoring* 17.7 (2010), pp. 825–847. DOI: 10.1002/stc.424.

[3] M. Davidian and D. M. Giltinan. “Nonlinear Models for Repeated Measurement Data: An Overview and Update”. In: *Journal of Agricultural, Biological, and Environmental Statistics* 8.4 (2003), pp. 387–419. DOI: 10.1198/1085711032697.

[4] H. T. Banks, Z. R. Kenz, and W. C. Thompson. "A review of selected techniques in inverse problem nonparametric probability distribution estimation". In: *Journal of Inverse and Ill-posed Problems* 20.4 (2012), pp. 429–460. DOI: 10.1515/jip-2012-0037.

[5] M. H. Faber. "On the Treatment of Uncertainties and Probabilities in Engineering Decision Analysis". In: *Journal of Offshore Mechanics and Arctic Engineering* 127.3 (2005), pp. 243–248. DOI: 10.1115/1.1951776.

[6] A. Der Kiureghian and O. Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009), pp. 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

[7] E. de Rocquigny and S. Cambier. "Inverse probabilistic modelling of the sources of uncertainty: A non-parametric simulated-likelihood method with application to an industrial turbine vibration assessment". In: *Inverse Problems in Science and Engineering* 17.7 (2009), pp. 937–959. DOI: 10.1080/17415970902916987.

[8] G. Celeux, A. Grimaud, Y. Lefèbvre, and E. de Rocquigny. "Identifying intrinsic variability in multivariate systems through linearized inverse methods". In: *Inverse Problems in Science and Engineering* 18.3 (2010), pp. 401–415. DOI: 10.1080/17415971003624330.

[9] P. Barbillon, G. Celeux, A. Grimaud, Y. Lefèbvre, and E. de Rocquigny. "Nonlinear methods for inverse statistical problems". In: *Computational Statistics & Data Analysis* 55.1 (2011), pp. 132–142. DOI: 10.1016/j.csda.2010.05.030.

[10] S. Fu, G. Celeux, N. Bousquet, and M. Couplet. "Bayesian Inference for Inverse Problems Occurring in Uncertainty Analysis". In: *International Journal for Uncertainty Quantification* 5.1 (2015), pp. 73–98. DOI: 10.1615/Int.J.UncertaintyQuantification.2014011073.

[11] I. Behmanesh, B. Moaveni, G. Lombaert, and C. Papadimitriou. "Hierarchical Bayesian model updating for structural identification". In: *Mechanical Systems and Signal Processing* 64–65 (2015), pp. 360–376. DOI: 10.1016/j.ymssp.2015.03.026.

[12] J. B. Nagel and B. Sudret. "A unified framework for multilevel uncertainty quantification in Bayesian inverse problems". In: *Probabilistic Engineering Mechanics* 43 (2016), pp. 68–84. DOI: 10.1016/j.probengmech.2015.09.007.

[13] D. Draper, D. P. Gaver, P. K. Goel, J. B. Greenhouse, L. V. Hedges, C. N. Morris, and C. M. Waternaux. *Combining Information: Statistical Issues and Opportunities for Research*. Washington, D.C., USA: Panel on Statistical Issues et al., 1992.

[14] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. 2nd ed. Springer Series in Statistics. New York: Springer, 2004. DOI: 10.1007/978-1-4757-4145-2.

[15] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds. *Handbook of Markov Chain Monte Carlo*. Handbooks of Modern Statistical Methods. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2011. DOI: 10.1201/b10905.

[16] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (1987), pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.

[17] J. B. Nagel and B. Sudret. "Probabilistic Inversion for Estimating the Variability of Material Properties: A Bayesian Multilevel Approach". In: *11th International Probabilistic Workshop (IPW11)*. Ed. by D. Novák and M. Vořechovský. Brno, Czech Republic: Litera, 2013, pp. 293–303. DOI: 10.3929/ethz-a-010034843.

[18] G. C. Ballesteros, P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos. "Bayesian Hierarchical Models for Uncertainty Quantification in Structural Dynamics". In: *2nd International Conference on Vulnerability and Risk Analysis and Management and 6th International Symposium on Uncertainty Modeling and Analysis (ICVRAM & ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 162, pp. 1615–1624. DOI: 10.1061/9780784413609.162.

[19] J. B. Nagel and B. Sudret. "Bayesian Multilevel Model Calibration for Inverse Problems Under Uncertainty with Perfect Data". In: *Journal of Aerospace Information Systems* 12.1 (2015), pp. 97–113. DOI: 10.2514/1.I010264.

[20] S. Särkkä. *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks. Cambridge, UK: Cambridge University Press, 2013. DOI: 10.1017/CBO9781139344203.

[21] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. New York: Springer, 2004. DOI: 10.1007/978-0-387-76371-2.

[22] R. M. Neal. "MCMC Using Hamiltonian Dynamics". In: *Handbook of Markov Chain Monte Carlo*. Ed. by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. Handbooks of Modern Statistical Methods. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2011. Chap. 5, pp. 113–162. DOI: 10.1201/b10905-6.

[23] J. M. Sanz-Serna. "Markov Chain Monte Carlo and Numerical Differential Equations". In: *Current Challenges in Stability Issues for Numerical Differential Equations*. Lecture Notes in Mathematics 2082. Cham, Switzerland: Springer International Publishing, 2014, pp. 39–88. DOI: 10.1007/978-3-319-01300-8_2.

[24] C. Gattringer and C. B. Lang. *Quantum Chromodynamics on the Lattice: An Introductory Presentation*. Lecture Notes in Physics 788. Springer-Verlag Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-01850-3.

[25] R. M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics 118. New York: Springer, 1996. DOI: 10.1007/978-1-4612-0745-0.

[26] B. Shahbaba, S. Lan, W. O. Johnson, and R. M. Neal. "Split Hamiltonian Monte Carlo". In: *Statistics and Computing* 24.3 (2014), pp. 339–349. DOI: 10.1007/s11222-012-9373-1.

[27] T. Papamarkou, A. Mira, and M. Girolami. "Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms". In: *Bayesian Analysis* 9.1 (2014), pp. 97–128. DOI: 10.1214/13-BA848.

[28] J. Sohl-Dickstein, M. Mudigonda, and M. DeWeese. "Hamiltonian Monte Carlo Without Detailed Balance". In: *JMLR Workshop and Conference Proceedings: 31st International Conference on Machine Learning (ICML 2014)* 32.1 (2014), pp. 719–726.

[29] C. M. Campos and J. Sanz-Serna. "Extra Chance Generalized Hybrid Monte Carlo". In: *Journal of Computational Physics* 281 (2015), pp. 365–374. DOI: 10.1016/j.jcp.2014.09.037.

[30] M. Girolami and B. Calderhead. "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2 (2011), pp. 123–214. DOI: 10.1111/j.1467-9868.2010.00765.x.

[31] M. Betancourt. "A General Metric for Riemannian Manifold Hamiltonian Monte Carlo". In: *Geometric Science of Information*. Ed. by F. Nielsen and F. Barbaresco. Lecture Notes in Computer Science 8085. Springer-Verlag Berlin Heidelberg, 2013, pp. 327–334. DOI: 10.1007/978-3-642-40020-9_35.

[32] A. Beskos, K. Kalogeropoulos, and E. Pazos. "Advanced MCMC methods for sampling on diffusion pathspace". In: *Stochastic Processes and their Applications* 123.4 (2013), pp. 1415–1453. DOI: 10.1016/j.spa.2012.12.001.

[33] A. H. Elsheikh, M. F. Wheeler, and I. Hoteit. "Hybrid nested sampling algorithm for Bayesian model selection applied to inverse subsurface flow problems". In: *Journal of Computational Physics* 258 (2014), pp. 319–337. DOI: 10.1016/j.jcp.2013.10.001.

[34] A. Kramer, B. Calderhead, and N. Radde. "Hamiltonian Monte Carlo methods for efficient parameter estimation in steady state dynamical systems". In: *BMC Bioinformatics* 15.1, 253 (2014), pp. 1–11. DOI: 10.1186/1471-2105-15-253.

[35] S. H. Cheung and J. L. Beck. "Bayesian Model Updating Using Hybrid Monte Carlo Simulation with Application to Structural Dynamic Models with Many Uncertain Parameters". In: *Journal of Engineering Mechanics* 135.4 (2009), pp. 243–255. DOI: 10.1061/(ASCE)0733-9399(2009)135:4(243).

[36] I. Boulkaibet, L. Mthembu, T. Marwala, M. I. Friswell, and S. Adhikari. "Finite element model updating using the shadow hybrid Monte Carlo technique". In: *Mechanical Systems and Signal Processing* 52–53 (2015), pp. 115–132. DOI: 10.1016/j.ymssp.2014.06.005.

[37] *Stan Modeling Language: User's Guide and Reference Manual*. 2.5.0. Stan Development Team. 2014.

[38] M. D. Hoffman and A. Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15 (2014), pp. 1593–1623.

[39] Y. Zhang and C. Sutton. "Semi-Separable Hamiltonian Monte Carlo for Inference in Bayesian Hierarchical Models". In: *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Morehouse Lane, New York, USA: Curran Associates, Inc., 2014, pp. 10–18.

[40] M. Betancourt and M. Girolami. "Hamiltonian Monte Carlo for Hierarchical Models". In: *Current Trends in Bayesian Methodology with Applications*. Ed. by S. K. Upadhyay, U. Singh, D. K. Dey, and A. Loganathan. Boca Raton, Florida, USA: CRC Press, 2015, pp. 79–101. DOI: 10.1201/b18502-5.

[41] O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. "A General Framework for the Parametrization of Hierarchical Models". In: *Statistical Science* 22.1 (2007), pp. 59–73. DOI: 10.1214/088342307000000014.

[42] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[43] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97.

[44] G. O. Roberts, A. Gelman, and W. R. Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *The Annals of Applied Probability* 7.1 (1997), pp. 110–120. DOI: 10.1214/aoap/1034625254.

[45] G. O. Roberts and J. S. Rosenthal. "Optimal scaling for various Metropolis-Hastings algorithms". In: *Statistical Science* 16.4 (2001), pp. 351–367. DOI: 10.1214/ss/1015346320.

[46] C. J. Geyer. "Practical Markov Chain Monte Carlo". In: *Statistical Science* 7.4 (1992), pp. 473–483. DOI: 10.1214/ss/1177011137.

[47] L. Tierney. "Markov Chains for Exploring Posterior Distributions". In: *The Annals of Statistics* 22.4 (1994), pp. 1701–1728. DOI: 10.1214/aos/1176325750.

[48] G. L. Jones. "On the Markov chain central limit theorem". In: *Probability Surveys* 1 (2004), pp. 299–320. DOI: 10.1214/154957804100000051.

[49] G. O. Roberts and J. S. Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability Surveys* 1 (2004), pp. 20–71. DOI: 10.1214/154957804100000024.

[50] A. Deriglazov. *Classical Mechanics: Hamiltonian and Lagrangian Formalism*. Springer-Verlag Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-14037-2.

[51] F. Schwabl. *Statistical Mechanics*. 2nd ed. Advanced Texts in Physics. Springer-Verlag Berlin Heidelberg, 2006. DOI: 10.1007/3-540-36217-7.

[52] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian Problems*. Applied Mathematics and Mathematical Computation 7. London, UK: Chapman & Hall/CRC, 1994.

[53] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics 14. Cambridge, UK: Cambridge University Press, 2004. DOI: 10.1017/CBO9780511614118.

[54] E. Hairer, C. Lubich, and G. Wanner. "Geometric numerical integration illustrated by the Störmer-Verlet method". In: *Acta Numerica* 12 (2003), pp. 399–450. DOI: 10.1017/S0962492902000144.

[55] S. Blanes, F. Casas, and J. Sanz-Serna. "Numerical Integrators for the Hybrid Monte Carlo Method". In: *SIAM Journal on Scientific Computing*. A 36.4 (2014), pp. 1556–1580. DOI: 10.1137/130932740.

[56] J. B. Nagel and B. Sudret. "A Bayesian Multilevel Approach to Optimally Estimate Material Properties". In: *2nd International Conference on Vulnerability and Risk Analysis and Management and 6th International Symposium on Uncertainty Modeling and Analysis (ICVRAM & ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 151, pp. 1504–1513. DOI: 10.1061/9780784413609.151.

[57] T. Bui-Thanh and M. Girolami. "Solving large-scale PDE-constrained Bayesian inverse problems with Riemann manifold Hamiltonian Monte Carlo". In: *Inverse Problems* 30.11, 114014 (2014), pp. 1–23. DOI: 10.1088/0266-5611/30/11/114014.

[58] B. Sudret and C. V. Mai. "Computing derivative-based global sensitivity measures using polynomial chaos expansions". In: *Reliability Engineering & System Safety* 134 (2015), pp. 241–250. DOI: 10.1016/j.ress.2014.07.009.

[59] A. Pakman and L. Paninski. "Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians". In: *Journal of Computational and Graphical Statistics* 23.2 (2014), pp. 518–542. DOI: 10.1080/10618600.2013.788448.

[60] S. Lan, B. Zhou, and B. Shahbaba. "Spherical Hamiltonian Monte Carlo for Constrained Target Distributions". In: *JMLR Workshop and Conference Proceedings: 31st International Conference on Machine Learning (ICML 2014)* 32.1 (2014), pp. 629–637.

[61] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, and A. Stuart. "Optimal tuning of the hybrid Monte Carlo algorithm". In: *Bernoulli* 19.5A (2013), pp. 1501–1534. DOI: 10.3150/12-BEJ414.

[62] M. Betancourt, S. Byrne, and M. Girolami. *Optimizing The Integrator Step Size for Hamiltonian Monte Carlo*. 2014. arXiv: 1411.6669 [stat.ME].

[63] E. Zio. *The Monte Carlo Simulation Method for System Reliability and Risk Analysis*. Springer Series in Reliability Engineering. London: Springer, 2013. DOI: 10.1007/978-1-4471-4588-2.

[64] P. B. Mackenzie. "An improved hybrid Monte Carlo method". In: *Physics Letters B* 226.3–4 (1989), pp. 369–371. DOI: 10.1016/0370-2693(89)91212-4.

[65] J. M. Flegal, M. Haran, and G. L. Jones. "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?" In: *Statistical Science* 23.2 (2008), pp. 250–260. DOI: 10.1214/08-STS257.

[66] J. M. Flegal and G. L. Jones. "Batch means and spectral variance estimators in Markov chain Monte Carlo". In: *The Annals of Statistics* 38.2 (2010), pp. 1034–1070. DOI: 10.1214/09-AOS735.

[67] B. D. Ripley. *Stochastic Simulation*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc., 1987. DOI: 10.1002/9780470316726.

[68] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal. "Markov Chain Monte Carlo in Practice: A Roundtable Discussion". In: *The American Statistician* 52.2 (1998), pp. 93–100. DOI: 10.1080/00031305.1998.10480547.

Chapter 6

Bayesian multilevel model calibration with perfect data

Original publication

J. B. Nagel and B. Sudret. "Bayesian Multilevel Model Calibration for Inverse Problems Under Uncertainty with Perfect Data". In: *Journal of Aerospace Information Systems* 12.1 (2015), pp. 97–113. DOI: 10.2514/1.I010264

Abstract

A probabilistic framework for Bayesian inference and uncertainty analysis is developed. It allows to address inverse problems in experimental situations where data is scarce and uncertainty is ubiquitous. The uncertainty characterization subproblem of the NASA Langley multidisciplinary uncertainty quantification challenge serves as the motivating application example. From responses of a computational model the goal is to learn about unknown model inputs that are subject to multiple types of uncertainty. This objective is interpreted and solved as Bayesian multilevel model calibration. The zero-noise or "perfect" data limit is investigated. Thereby the likelihood function is defined as a solution to forward uncertainty propagation. Posterior explorations are based on suitable Markov chain Monte Carlo algorithms and stochastic likelihood simulations. An unforeseen finding in this context is that the posterior distribution can only be sampled with a certain degree of fidelity. Partial data augmentation is introduced as a means to improve the error statistics of likelihood estimations and the fidelity of posterior computations.

6.1 Introduction

The NASA Langley multidisciplinary uncertainty quantification challenge has raised contemporary open questions to uncertainty quantification (UQ) [1, 2]. Altogether it consists of five subproblems that deal with uncertainty characterization, sensitivity analysis, uncertainty propagation, extreme case analysis and robust design. These problems originate from a specific aerospace application which is part of greater efforts to reduce the rate of fatal loss-of-control accidents [3–5]. An abstract and widely discipline-independent problem formulation prompts researchers and practitioners from various fields in academia and industry to devise generic solutions to the problems. A dynamically scaled, free-flight model of a remotely piloted, twin-turbine powered transport aircraft is the physical system under consideration. It serves as a prototyping and experimentation testbed for flight control in adverse situations, e.g. under structural damage or component failure. Parameter uncertainties of this subscaled model reflect the uncertainties in aerodynamic conditions and losses in control effectiveness.

In this contribution we address the uncertainty characterization subproblem of the challenge posed. With given responses of a computational model the challenge is to learn about the unknown inputs that parametrize the flying conditions. Throughout the experiments data are collected while model inputs are subject to epistemic uncertainty and aleatory variability [6, 7]. Inference therefore focuses on physically fixed model parameters as well as on so-called hyperparameters. The latter determine the distribution of such model inputs that are variable during experimentation. We approach the problem from a Bayesian perspective to statistical inversion and uncertainty quantification [8–10]. While classical Bayesian inversion allows to estimate constant model parameters, the additional identification of hyperparameters requires hierarchical modeling approaches.

Hierarchical models were mainly developed in biological statistics [11, 12] and they are only slowly being adopted within the engineering community [13–15].

The goal of this paper is the development of a framework along with computational tools for attacking inverse problems under aleatory and epistemic parameter uncertainty. We combine classical inversion and hierarchical modeling into a Bayesian multilevel framework that allows to tackle the general class of problems that the uncertainty characterization subproblem typifies. Within a probabilistic setting this eventually allows for an elegant formulation and efficient numerical solution. The foundations of inverse modeling in conjunction with "perfect" data, i.e. in the zero-noise limit, and parameter uncertainty are laid. Randomness in the data is then solely attributed to a probability model of the input arguments of a computational "blackbox" solver. The likelihood is formulated as a solution to uncertainty propagation. Since this renders its evaluation analytically intractable, statistical estimators based on the Monte Carlo method and kernel density estimation are proposed. In this context the induced type of posterior approximation is investigated. Heuristic ways of tuning free algorithmic parameters, e.g. the kernel bandwidth, are presented. Partial data augmentation is proposed in order to improve likelihood estimations through automatic kernel bandwidth selection and to enhance the fidelity of the posterior.

The manuscript is organized as follows. In Section 6.2 a generic Bayesian multilevel framework for inversion under uncertainty will be initially formulated. For the solution of the NASA Langley UQ challenge we will devise a statistical model involving "perfect" data in Section 6.3. Computational key challenges posed by Bayesian inference in the present context will be discussed in Section 6.4. The challenge problem will be cast as multilevel inversion in Section 6.5 and in the subsequent Section 6.6 our results will be presented. In Section 6.7 data augmentation will be utilized in order to ensure a sufficient degree of algorithmic efficiency and posterior fidelity. We will conclude in Section 6.8 where the gathered experience from solving the NASA UQ challenge problem is summarized.

6.2 Bayesian multilevel modeling

Due to the lack of a universally accepted terminology, we define a *multilevel* or *hierarchical model* as "an assembly of submodels at different levels of a hierarchy". The hierarchical structure can be constituted by stochastic dependencies and deterministic maps between the quantities involved. According to that definition multilevel modeling forms sort of an overarching theme in modern cross-disciplinary statistics. In the last two decades it has been extensively studied from a frequentist [11, 12] and a Bayesian [16, 17] point of view. Adopting the latter paradigm, prior elicitation [18, 19] and posterior computation [20, 21] are delicate issues that have been discussed in the statistical literature. Applications of multilevel modeling encompass probabilistic inversion [22, 23] and optimal combination of information [24, 25]. Based on a probabilistic representation of both epistemic uncertainty and aleatory variability, Bayesian multilevel modeling establishes a natural framework for solving complex inverse problems under uncertainty. Inference can be accomplished by transforming, conditioning and marginalizing probability distributions.

6.2.1 Uncertainty and variability

A *forward model* $\mathcal{M}\colon (\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \mapsto \tilde{\boldsymbol{y}}$ represents the system or phenomenon under consideration. It formally maps model inputs $(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \in \mathcal{D}_{\boldsymbol{m}} \times \mathcal{D}_{\boldsymbol{x}} \times \mathcal{D}_{\boldsymbol{\zeta}} \times \mathcal{D}_{\boldsymbol{d}}$ to outputs $\tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \in \mathcal{D}_{\tilde{\boldsymbol{y}}} \subset \mathbb{R}^d$. When carrying out a number of experiments the variability of measured forward model responses can be attributed to models of input uncertainty. There are fixed yet unknown model parameters $\boldsymbol{m} \in \mathcal{D}_{\boldsymbol{m}} \subset \mathbb{R}^p$, model inputs $\boldsymbol{\zeta} \in \mathcal{D}_{\boldsymbol{\zeta}} \subset \mathbb{R}^r$ with perfectly known aleatory variability, input variables $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}} \subset \mathbb{R}^q$ with imperfectly known aleatory variability, and experimental conditions $\boldsymbol{d} \in \mathcal{D}_{\boldsymbol{d}} \subset \mathbb{R}^s$ that are entirely known.

With respect to a number of $i = 1, \ldots, n$ experiments, forward model inputs are represented as deterministic or stochastic objects within the Bayesian multilevel framework. Throughout the experiments data is acquired under known but possibly different experimental conditions $\boldsymbol{d}_i$. These model inputs $\boldsymbol{d}_i$ are therefore deterministically represented. Fixed albeit unknown model parameters $\boldsymbol{m}$ are assumed to be constant over the experiments. In Bayesian fashion they are represented as random variables $\boldsymbol{M} \sim \pi_{\boldsymbol{M}}(\boldsymbol{m})$, where the Bayesian prior distribution $\pi_{\boldsymbol{M}}(\boldsymbol{m})$ accounts for a subjective degree of belief or prior knowledge about their true values. This is the Bayesian conception of *epistemic uncertainty*.

Over the number of experiments varying model inputs $\boldsymbol{\zeta}$ take on unknown experiment-specific realizations $\boldsymbol{\zeta}_i$ of conditionally independent random variables $(\boldsymbol{Z}_i|\boldsymbol{\theta}_{\boldsymbol{Z}}) \sim f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i|\boldsymbol{\theta}_{\boldsymbol{Z}})$. The conditional distribution $f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i|\boldsymbol{\theta}_{\boldsymbol{Z}})$ with known hyperparameters $\boldsymbol{\theta}_{\boldsymbol{Z}}$ states a subjective degree of belief or prior knowledge about the individual realizations $\boldsymbol{\zeta}_i$. This is a Bayesian notion of *aleatory variability*. Similarly model inputs $\boldsymbol{x}$ are subject to variability and take on unknown experiment-specific realizations $\boldsymbol{x}_i$ of conditionally independent

random variables $(\boldsymbol{X}_i|\boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})$. The hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ determine this variability throughout the experiments and are fixed but unknown. In turn they are modeled as random variables $\boldsymbol{\Theta}_{\boldsymbol{X}} \sim \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$, where the distribution $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ quantifies an a priori degree of plausibility. Random variables $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \sim \int (\prod_{i=1}^{n} f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}}))\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})\, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}$ embody the prior knowledge about the experiment-specific realizations $\boldsymbol{x}_i$.

Summarizing, marginal distributions $\pi_{\boldsymbol{M}}(\boldsymbol{m})$ and $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ represent *parametric prior knowledge* about the true values of the model parameters $\boldsymbol{m}$ and the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$, whereas conditional distributions $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})$ and $f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i|\boldsymbol{\theta}_{\boldsymbol{Z}})$ encapsulate *structural prior knowledge* about the problem, i.e. information about experiment-specific $\boldsymbol{x}_i$ and $\boldsymbol{\zeta}_i$.

6.2.2 Statistical data model

An integral constituent of many statistical approaches to inverse problems is a residual model. Real observations $\boldsymbol{y}_i$ often deviate from model predictions $\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ even if forward model inputs were known with certainty. This discrepancy, which is due to measurement errors, numerical approximations and model inadequacies, is often accounted for by a statistical data model $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i$. Prediction errors $\boldsymbol{\varepsilon}_i$ are assumed to be realizations of random variables $\boldsymbol{E}_i \sim f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i)$, e.g. with normal distributions $f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i) = \mathcal{N}(\boldsymbol{\varepsilon}_i|\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ and experiment-specific, symmetric and positive-semidefinite covariance matrices $\boldsymbol{\Sigma}_i$. It quantifies a degree of imperfection of the forward model and experimental apparatus. Hence observations are viewed as realizations $\boldsymbol{y}_i$ of random variables $(\boldsymbol{Y}_i|\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i)$ with distributions $f(\boldsymbol{y}_i|\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i) = f_{\boldsymbol{E}_i}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i))$. The overall model formulated thus far can be summarized as

$$(\boldsymbol{Y}_i|\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i) \sim f_{\boldsymbol{E}_i}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)\big), \tag{6.1a}$$

$$\boldsymbol{M} \sim \pi_{\boldsymbol{M}}(\boldsymbol{m}), \tag{6.1b}$$

$$(\boldsymbol{X}_i|\boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}}), \tag{6.1c}$$

$$\boldsymbol{\Theta}_{\boldsymbol{X}} \sim \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}), \tag{6.1d}$$

$$(\boldsymbol{Z}_i|\boldsymbol{\theta}_{\boldsymbol{Z}}) \sim f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i|\boldsymbol{\theta}_{\boldsymbol{Z}}). \tag{6.1e}$$

This model is composed of conditional probabilistic and deterministic relations between the quantities involved. As per our previous definition it is a generic Bayesian multilevel model. An intuitive model representation is provided by a directed acyclic graph (DAG) [26, 27], such as shown in Fig. 6.1. In Section 6.3 we will devise a model for analyzing "perfect" observations $\tilde{\boldsymbol{y}}_i$ instead of the "imperfect" ones $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i$. Unlike the latter, the randomness of the former is exclusively attributed to forward model input variability as discussed in Section 6.2.1. Unless stated or denoted otherwise random variables in Eq. (6.1) are assumed to be (conditionally) independent. This defines a joint overall probability density of all probabilistic quantities. By conditioning and marginalizing this overall density at one's convenience, one can derive meaningful probability densities. For inferential purposes these are certain prior and posterior distributions that we will explain in the following.

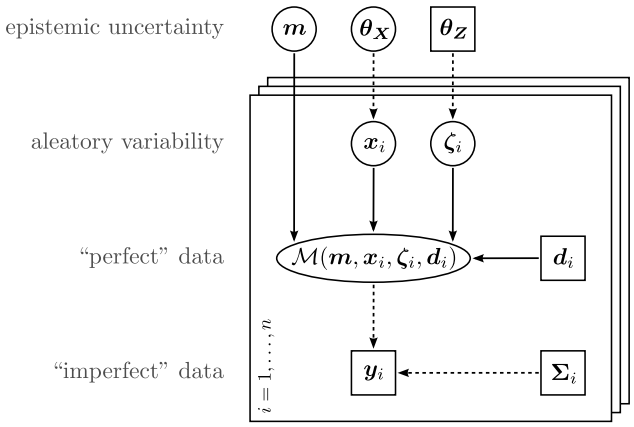


Figure 6.1: DAG of the generic multilevel model. Vertices symbolize unknown (○) or known (□) quantities and directed edges represent their deterministic (⟶) or probabilistic (⇢) relations. Model parameters $\boldsymbol{m}$ are constant over $i = 1, \dots, n$ experiments. The variability of experiment-specific realizations $\boldsymbol{x}_i$ and $\boldsymbol{\zeta}_i$ is determined by unknown or known hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ and $\boldsymbol{\theta}_{\boldsymbol{Z}}$, respectively. Data can be interpreted as "perfect" $\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ or "imperfect" observations $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i$ with $f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i) = \mathcal{N}(\boldsymbol{\varepsilon}_i|\boldsymbol{0}, \boldsymbol{\Sigma}_i)$.

6.2.3 Inference in multilevel models

In what follows $\langle \boldsymbol{q}_i \rangle$ denotes a tuple $\langle \boldsymbol{q}_i \rangle_{1 \le i \le n} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_n)$. Conditioned on the priorly known hyperparameters $\boldsymbol{\theta}_{\boldsymbol{Z}}$, the joint prior distribution of the unknowns $(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}})$ is given as

$$\pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \boldsymbol{\theta}_{\boldsymbol{Z}}\big) = \left(\prod_{i=1}^{n} f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \right) \left(\prod_{i=1}^{n} f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i | \boldsymbol{\theta}_{\boldsymbol{Z}}) \right) \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})\, \pi_{\boldsymbol{M}}(\boldsymbol{m}). \tag{6.2}$$

It summarizes the available parametric and structural prior knowledge. The joint posterior distribution of the unknowns $(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}})$ is obtained by further conditioning the prior Eq. (6.2) on the data $\langle \boldsymbol{y}_i \rangle$. By virtue of Bayes' law this posterior is up to a scale factor found as

$$\pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{Z}}\big) \propto \left(\prod_{i=1}^{n} f_{\boldsymbol{E}_i}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)\big) \right) \pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \boldsymbol{\theta}_{\boldsymbol{Z}}\big). \tag{6.3}$$

The posterior degree of plausibility about the *quantities of interest* (QoI) can be extracted by marginalizing the posterior Eq. (6.3) over parameters considered *nuisance* [28, 29]. Provided $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ are QoI and $(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$ are nuisance parameters, the correspondingly marginalized posterior is

$$\pi\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{Z}}\big) = \int_{\mathcal{D}_{\boldsymbol{x}}^n} \int_{\mathcal{D}_{\boldsymbol{\zeta}}^n} \pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{Z}}\big)\, \mathrm{d}\langle \boldsymbol{x}_i \rangle\, \mathrm{d}\langle \boldsymbol{\zeta}_i \rangle, \tag{6.4}$$

where $\mathrm{d}\langle \boldsymbol{x}_i \rangle = \mathrm{d}\boldsymbol{x}_1 \dots \mathrm{d}\boldsymbol{x}_n$ and $\mathrm{d}\langle \boldsymbol{\zeta}_i \rangle = \mathrm{d}\boldsymbol{\zeta}_1 \dots \mathrm{d}\boldsymbol{\zeta}_n$. Summarized the genuinely unique approach to Bayesian inference in multilevel models is to construct the posterior of the QoI $(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$ by conditioning on the knowns $(\langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{Z}})$ and subsequently marginalizing out nuisance $(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$.

6.2.4 Marginalized formulation

Equivalently one could solve a marginal formulation of the multilevel calibration problem, with a marginal prior $\pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \pi_{\boldsymbol{M}}(\boldsymbol{m})\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ and a marginalized version of the likelihood $\mathcal{L}(\langle \boldsymbol{y}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}})$ [30, 31]. One therefore factorizes the marginalized posterior Eq. (6.4) as

$$\pi\big(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{Z}}\big) \propto \left(\prod_{i=1}^{n} f(\boldsymbol{y}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}}) \right) \pi_{\boldsymbol{M}}(\boldsymbol{m})\, \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}), \tag{6.5}$$

where $f(\boldsymbol{y}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}})$ is the marginalized density of the observation $\boldsymbol{y}_i$. The aleatory variables $(\boldsymbol{x}_i, \boldsymbol{\zeta}_i)$ have been eliminated based on the integration

$$f(\boldsymbol{y}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}}) = \int_{\mathcal{D}_{\boldsymbol{x}}} \int_{\mathcal{D}_{\boldsymbol{\zeta}}} f_{\boldsymbol{E}_i}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)\big)\, f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})\, f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i | \boldsymbol{\theta}_{\boldsymbol{Z}})\, \mathrm{d}\boldsymbol{x}_i\, \mathrm{d}\boldsymbol{\zeta}_i. \tag{6.6}$$

When defining the *marginalized* or *integrated likelihood* as $\mathcal{L}(\langle \boldsymbol{y}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}})$, the marginal posterior Eq. (6.5) simply writes as $\pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}} | \langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta}_{\boldsymbol{Z}}) \propto \mathcal{L}(\langle \boldsymbol{y}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}})\, \pi(\boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}})$. Directly computing the marginal posterior Eq. (6.4) involves a lower-dimensional parameter space compared to computing the joint posterior Eq. (6.3). However, it requires the computation of the integrals in Eq. (6.6).

6.3 "Perfect" data model

In Section 6.2.2 the residual model was introduced as a representation of the "imperfection" of the forward solver and measurement device. As a consequence, in Eq. (6.1a) observations were regarded as realizations $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i$ of random variables $(\boldsymbol{Y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i)$ conditioned on direct forward model inputs. The residual assumptions that led to Eq. (6.1a) had equipped the data space $\mathcal{D}_{\tilde{\boldsymbol{y}}}$ with a probability model. We introduce the term "imperfect" for this statistical data model in order to distinguish it from the following.

In this paper we are interested in the experimental situation where $\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ can be directly observed, e.g. due to noise-free measurements and a "sufficiently accurate" forward simulator. Hereinafter we will refer to this limiting case as to involve "perfect" data. Not being premised on a residual model, the data are viewed as realizations $\tilde{\boldsymbol{y}}_i$ of random variables $(\tilde{\boldsymbol{Y}}_i | \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}})$ conditioned on the "highest-level" quantities.

Provided an appropriate probability model $f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})$ of those random variables, a Bayesian multilevel model for "perfect" data can be written as

$$(\tilde{\boldsymbol{Y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}}) \sim f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}}), \tag{6.7a}$$

$$(\boldsymbol{M},\boldsymbol{\Theta}_{\boldsymbol{X}}) \sim \pi(\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}}) = \pi_{\boldsymbol{M}}(\boldsymbol{m})\,\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{6.7b}$$

As before prior knowledge about the unknowns $(\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}})$ is embodied in Eq. (6.7b). As a function of the unknowns $(\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}})$, with the density Eq. (6.7a) one can formulate a residual-free version of the likelihood

$$\mathcal{L}\big(\langle\tilde{\boldsymbol{y}}_i\rangle|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}}\big) = \prod_{i=1}^{n} f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}}). \tag{6.8}$$

For reasons that will be discussed below, we call Eq. (6.8) the *transformed likelihood*. As usual Bayesian analysis proceeds by conditioning the prior on the acquired data $\langle\tilde{\boldsymbol{y}}_i\rangle$. The posterior follows as

$$\pi\big(\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}}\,|\,\langle\tilde{\boldsymbol{y}}_i\rangle,\boldsymbol{\theta}_{\boldsymbol{Z}}\big) \propto \mathcal{L}\big(\langle\tilde{\boldsymbol{y}}_i\rangle|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}}\big)\,\pi(\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{6.9}$$

Note that the notation of Eq. (6.9) is reminiscent of classical Bayesian inversion. Indeed the multilevel character of the problem manifests in the likelihood function Eq. (6.8) that we will now specify in greater detail.

6.3.1 Forward uncertainty propagation

Let $\boldsymbol{X}_i \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})$ and $\boldsymbol{Z}_i \sim f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i|\boldsymbol{\theta}_{\boldsymbol{Z}})$ be the aleatory variables in Eqs. (6.1c) and (6.1e) that have independent marginal densities for given hyperparameters $(\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})$. Now consider the map $\mathcal{M}_{\boldsymbol{m},\boldsymbol{d}_i}\colon(\boldsymbol{x}_i,\boldsymbol{\zeta}_i)\mapsto\tilde{\boldsymbol{y}}_i=\mathcal{M}(\boldsymbol{m},\boldsymbol{x}_i,\boldsymbol{\zeta}_i,\boldsymbol{d}_i)$ that the forward model defines for fixed inputs $(\boldsymbol{m},\boldsymbol{d}_i)$. For given $(\boldsymbol{m},\boldsymbol{d}_i,\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})$ the random variables in Eq. (6.7a) are constructed by forward propagation of the aleatory input uncertainties through the function $\mathcal{M}_{\boldsymbol{m},\boldsymbol{d}_i}(\boldsymbol{x}_i,\boldsymbol{\zeta}_i)$ into an output uncertainty. Provided the existence of a corresponding density, the transformed random variables $(\tilde{\boldsymbol{Y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})=\mathcal{M}_{\boldsymbol{m},\boldsymbol{d}_i}(\boldsymbol{X}_i,\boldsymbol{Z}_i)$ follow

$$f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}}) = \int_{\mathcal{D}_{\boldsymbol{x}}}\int_{\mathcal{D}_{\boldsymbol{\zeta}}} \delta\big(\tilde{\boldsymbol{y}}_i-\mathcal{M}_{\boldsymbol{m},\boldsymbol{d}_i}(\boldsymbol{x}_i,\boldsymbol{\zeta}_i)\big)\,f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{X}})\,f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i|\boldsymbol{\theta}_{\boldsymbol{Z}})\,\mathrm{d}\boldsymbol{x}_i\,\mathrm{d}\boldsymbol{\zeta}_i. \tag{6.10}$$

Here δ denotes the Dirac delta distribution. At this point it is to be noted that epistemic uncertainties are not propagated through the forward model. The transformed density Eq. (6.10) establishes the proper probability model in the space of data $\mathcal{D}_{\tilde{\boldsymbol{y}}}$ and thereby defines the transformed likelihood function Eq. (6.8). More formally one can derive Eq. (6.10) as the density function of a push-forward measure [32].

6.3.2 Relation between "imperfect" and "perfect" data

It is interesting to examine the relation between the "perfect" and the "imperfect" data model. The full multilevel model Eq. (6.1) was defined by a cascade of random variables that were conditioned on one another. Constructing the marginal model $(\boldsymbol{Y}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})$ in Eq. (6.6) was then based on the marginalization of aleatory variables, whereas the mechanism to formulate the "perfect" data model $(\tilde{\boldsymbol{Y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})$ in Eq. (6.10) was their propagation. Those two operations are related by writing $(\boldsymbol{Y}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})=(\tilde{\boldsymbol{Y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})+\boldsymbol{E}_i$ as the sum of independent random variables. The convolution integral

$$f(\boldsymbol{y}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}}) = \int_{\mathcal{D}_{\tilde{\boldsymbol{y}}}} f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})\,f_{\boldsymbol{E}_i}(\boldsymbol{y}_i-\tilde{\boldsymbol{y}}_i)\,\mathrm{d}\tilde{\boldsymbol{y}}_i \tag{6.11}$$

then establishes the relation between the distributions $f(\boldsymbol{y}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})$ and $f(\tilde{\boldsymbol{y}}_i|\boldsymbol{m},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{\theta}_{\boldsymbol{Z}})$ of "imperfect" and "perfect" data, respectively. Of course, when plugging Eq. (6.10) in Eq. (6.11) one easily re-derives Eq. (6.6).

For finite measurement uncertainty $\|\boldsymbol{\Sigma}_i\|>0$ the two models involving "imperfect" and "perfect" data describe distinct experimental situations. However, in the limiting case $\|\boldsymbol{\Sigma}_i\|\to 0$ with $\mathbb{E}[((\boldsymbol{Y}_i|\boldsymbol{m},\boldsymbol{x}_i,\boldsymbol{\zeta}_i)-\tilde{\boldsymbol{y}}_i)^2]\to 0$ and $f_{\boldsymbol{E}_i}\to\delta$ the marginalized likelihood approaches the transformed likelihood. Naturally this meets one's expectations. In classical Bayesian inversion the small-noise limit implies that the posterior of the unobservables shrinks to exact solutions of the inverse problem, i.e. posterior consistency [33, 34]. In multilevel inversion, however, the zero-noise limit leads to convergence of the posterior Eq. (6.5) to Eq. (6.9).

6.3.3 Kernel density estimation

Since the transformed likelihood Eq. (6.8) will be rarely available in analytical form, one has commonly to rely on numerical approximations. A possible approach is to simulate the response density Eq. (6.10) by Monte Carlo (MC) sampling and kernel density estimation (KDE) [35, 36] and evaluate the likelihood accordingly [37, 38].

In the d-variate case, given a sample $(\tilde{\boldsymbol{y}}^{(1)}, \dots, \tilde{\boldsymbol{y}}^{(K)})$ from some distribution with an unknown density $f(\tilde{\boldsymbol{y}})$, a kernel smoothing (KS) estimate of this density is given as $\hat{f}(\tilde{\boldsymbol{y}}) = K^{-1} \sum_{k=1}^{K} \mathcal{K}_{\boldsymbol{H}}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{y}}^{(k)})$. The scaled kernel $\mathcal{K}_{\boldsymbol{H}}(\tilde{\boldsymbol{y}}) = |\boldsymbol{H}|^{-1/2} \mathcal{K}(\boldsymbol{H}^{-1/2}\tilde{\boldsymbol{y}})$ is defined by a kernel function $\mathcal{K}$ and a symmetric and positive-definite bandwidth matrix $\boldsymbol{H}$. Common types of bandwidth matrices are multiples of the identity matrix $\boldsymbol{H} = h^2 \mathbf{1}_d$ for $h > 0$ or diagonal matrices $\boldsymbol{H} = \operatorname{diag}(h_1^2, \dots, h_d^2)$ with $h_1, \dots, h_d > 0$. According to certain criteria and assumptions, "optimal" bandwidths are commonly selected to prevent over- and undersmoothing. This amounts to a classical trade-off between the bias and the variance of the estimation. The mean squared error (MSE) of a KDE $\hat{f}(\tilde{\boldsymbol{y}})$ at a point $\tilde{\boldsymbol{y}}$ can be decomposed into its variance and the square of its bias

$$\operatorname{MSE}\Big[\hat{f}(\tilde{\boldsymbol{y}})\Big] = \mathbb{E}\Big[\big(\hat{f}(\tilde{\boldsymbol{y}}) - f(\tilde{\boldsymbol{y}})\big)^2\Big] = \operatorname{Var}\Big[\hat{f}(\tilde{\boldsymbol{y}})\Big] + \operatorname{Bias}^2\Big[\hat{f}(\tilde{\boldsymbol{y}})\Big]. \tag{6.12}$$

An optimal bandwidth normally strives to minimize the mean integrated squared error, i.e. when Eq. (6.12) is integrated over $\tilde{\boldsymbol{y}}$, by balancing the variance against the bias. The KDE of a univariate density similar to a Gaussian with kernels of the same type, can be based on Silverman's normal reference rule [35], i.e. $h = (4/3K)^{1/5}\, \hat{\sigma}$. Here $\hat{\sigma}$ is the standard deviation of the sample $(\tilde{y}^{(1)}, \dots, \tilde{y}^{(K)})$.

Based on MC and KDE techniques one can estimate the target density Eq. (6.10). On top of that we propose to estimate the transformed likelihood Eq. (6.8) through

$$\hat{\mathcal{L}}_{\mathrm{KS}}\big(\langle \tilde{\boldsymbol{y}}_i \rangle \,|\, \boldsymbol{m}, \boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}}\big) = \prod_{i=1}^{n} \left(\frac{1}{K} \sum_{k=1}^{K} \mathcal{K}_{\boldsymbol{H}} \left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_i^{(k)} \right) \right), \quad \text{with} \quad \begin{cases} \boldsymbol{x}^{(k)} \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}^{(k)} | \boldsymbol{\theta}_{\boldsymbol{X}}), \\ \boldsymbol{\zeta}^{(k)} \sim f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}^{(k)} | \boldsymbol{\theta}_{\boldsymbol{Z}}), \\ \tilde{\boldsymbol{y}}_i^{(k)} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}^{(k)}, \boldsymbol{\zeta}^{(k)}, \boldsymbol{d}_i). \end{cases} \tag{6.13}$$

For $k = 1, \dots, K$ forward model responses $\tilde{\boldsymbol{y}}_i^{(k)} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}^{(k)}, \boldsymbol{\zeta}^{(k)}, \boldsymbol{d}_i)$ are computed for given arguments $(\boldsymbol{m}, \boldsymbol{d}_i)$ and for further inputs $\boldsymbol{x}^{(k)}$ and $\boldsymbol{\zeta}^{(k)}$ that are randomly sampled from their parent distributions. These distributions $f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}^{(k)}|\boldsymbol{\theta}_{\boldsymbol{X}})$ and $f_{\boldsymbol{Z}|\boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}^{(k)}|\boldsymbol{\theta}_{\boldsymbol{Z}})$ are defined by values of the hyperparameters $(\boldsymbol{\theta}_{\boldsymbol{X}}, \boldsymbol{\theta}_{\boldsymbol{Z}})$, respectively.

With Eq. (6.13) we have a nonparametric statistical estimator of the likelihood Eq. (6.8) at hand. However, this estimator is accompanied by additional free parameters, i.e. the type of kernel function $\mathcal{K}$, the number of samples K and the bandwidth $\boldsymbol{H}$. As it will be discussed in the following, the application of classical criteria, that usually assist in adjusting free parameters, is questionable in the context of posterior computation via Markov chain Monte Carlo.

6.4 Bayesian computations

More often than not Bayesian posterior distributions do not have analytic closed-form solutions. Nevertheless one can explore posteriors through Markov chain Monte Carlo (MCMC) sampling techniques [39]. The principle of MCMC is to realize an ergodic Markov chain over the prior support whose invariant distribution equals the posterior. Let $\pi_0(\boldsymbol{q})$ be the prior and $\pi_1(\boldsymbol{q}) \propto \mathcal{L}(\boldsymbol{q})\, \pi_0(\boldsymbol{q})$ the posterior of an unknown QoI $\boldsymbol{q}$. The Markov kernel $\mathcal{K}$ defines the density $\mathcal{K}(\boldsymbol{q}^{(t)}, \boldsymbol{q}^{(t+1)})$ of the transition probability from $\boldsymbol{q}^{(t)}$ to $\boldsymbol{q}^{(t+1)}$, i.e. the state of the chain at times t and $t+1$. The posterior $\pi_1(\boldsymbol{q})$ is said to be an invariant distribution of the Markov chain if $\pi_1(\boldsymbol{q}^{(t+1)}) = \int \pi_1(\boldsymbol{q}^{(t)})\, \mathcal{K}(\boldsymbol{q}^{(t)}, \boldsymbol{q}^{(t+1)})\, \mathrm{d}\boldsymbol{q}^{(t)}$. This is abbreviated as $\pi_1 = \pi_1 \mathcal{K}$. Detailed balance, i.e. time reversibility $\pi_1(\boldsymbol{q}^{(t)})\, \mathcal{K}(\boldsymbol{q}^{(t)}, \boldsymbol{q}^{(t+1)}) = \pi_1(\boldsymbol{q}^{(t+1)})\, \mathcal{K}(\boldsymbol{q}^{(t+1)}, \boldsymbol{q}^{(t)})$, is a sufficient condition for the Markov chain to leave the posterior $\pi_1(\boldsymbol{q})$ invariant. It normally serves as the guiding principle for the construction of a Markov chain appropriate for posterior exploration. The Metropolis-Hastings (MH) algorithm establishes a prototypical class of MCMC techniques that relies on this principle [40, 41].

6.4.1 The Metropolis-Hastings algorithm

Initialized at $\boldsymbol{q}^{(0)}$ the MH algorithm generates a Markov chain with steady-state distribution $\pi_1(\boldsymbol{q})$ by iteratively applying the Markov chain transition kernel as follows. For a current state $\boldsymbol{q}^{(t)}$ of the Markov chain

a candidate state $\boldsymbol{q}^{(\star)} \sim P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})$ is sampled from a proposal distribution $P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})$. The proposal state becomes accepted, i.e. $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(\star)}$, with probability

$$\alpha\left(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)}\right) = \min\left(1, \frac{\pi_1(\boldsymbol{q}^{(\star)})\, P(\boldsymbol{q}^{(t)}|\boldsymbol{q}^{(\star)})}{\pi_1(\boldsymbol{q}^{(t)})\, P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})}\right). \tag{6.14}$$

Otherwise it is rejected, i.e. $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}$. The MH transition kernel defined this way satisfies detailed balance. Note that the MH correction Eq. (6.14) requires the computation of posterior ratios, hence only unscaled posterior densities have to be evaluated. Classical random walk Metropolis sampling is based on local proposals, e.g. sampling candidate states from a Gaussian $\boldsymbol{q}^{(\star)} \sim \mathcal{N}(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{q}})$ with mean $\boldsymbol{q}^{(t)}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{q}}$. Independence MH samplers are based on nonlocal proposals, e.g. sampling candidate states from the prior $\boldsymbol{q}^{(\star)} \sim \pi_0(\boldsymbol{q}^{(\star)})$ or from some suitable approximation of the posterior $\boldsymbol{q}^{(\star)} \sim \hat{\pi}_1(\boldsymbol{q}^{(\star)})$.

6.4.2 Key challenges

Typically MCMC sampling calls for a high number of forward model runs for likelihood evaluations in Eq. (6.14). Besides that, the degree as to which MCMC samples are autocorrelated governs their quality as posterior representatives. The design and efficient tuning of MCMC algorithms therefore aims at optimizing the mixing properties, i.e. the speed of convergence of the Markov chain towards its equilibrium distribution. This is a challenging and highly problem-dependent task. MCMC methods demand careful convergence diagnostics, i.e. the assessment of when the Markov chain has reached its target distribution and has lost the dependency on its initialization [42, 43]. Moreover MCMC suffers from difficulties in exploring high-dimensional parameter spaces and multimodal posteriors. Multilevel model calibration poses further multilevel-specific MCMC burdens. Sampling the posterior Eq. (6.9) imposes estimations of the likelihood Eq. (6.8) that is analytically intractable [44].

6.4.3 Posterior fidelity

Due to Bayes' law, deterministic closed-form approximations $\bar{\mathcal{L}}$ of the likelihood $\mathcal{L}$ directly induce approximations on the level of the posterior $\bar{\pi}_1(\boldsymbol{q}) \propto \bar{\mathcal{L}}(\boldsymbol{q})\, \pi_0(\boldsymbol{q})$. However, if the posterior is explored by means of MCMC and calls to the likelihood function $\mathcal{L}$ are replaced by calls to a statistical estimator $\hat{\mathcal{L}}$, then a modification is introduced on the level of the Markov chain transition kernel [45, 46]. There is no reason to expect that the modified MH transition kernel with the "randomized" version of Eq. (6.14), leaves the posterior invariant, i.e. in general $\pi_1 \neq \pi_1\hat{\mathcal{K}}$. Consequentially there arises the question of whether an equilibrium distribution $\hat{\pi}_1 = \hat{\pi}_1\hat{\mathcal{K}}$ actually exists, and in the event of that it does, as to what extent the induced distribution $\hat{\pi}_1$ is in congruence with the true posterior π_1. In order to pay tribute to these issues we introduce the term *posterior fidelity* as a qualitative measure of the similarity between $\hat{\pi}_1$ and π_1.

Moreover there is the practical question of how free algorithmic parameters, e.g. the number of response samples K and the kernel bandwidth $\boldsymbol{H}$, can be set in order to provide a convenient trade-off between posterior fidelity and computational feasibility, i.e. an "optimal" parameter tuning. We suppose that it is indeed possible to define certain criteria that parameter tuning can be based on. Even though this is beyond the scope of this paper, we have some preliminary comments. As done below, when postulating the existence and uniqueness of an equilibrium distribution, one can further argue that "small" changes in the transition kernel only cause "small" changes in the distribution. Given the current state $\boldsymbol{q}^{(t)}$ of the Markov chain, that is assumed to be ergodic, in the MH correction Eq. (6.14) a certain "random" decision is made whether to approve or to refuse a candidate state $\boldsymbol{q}^{(\star)}$. This binary decision follows the computation of the posterior ratio $\pi_1(\boldsymbol{q}^{(\star)})/\pi_1(\boldsymbol{q}^{(t)})$. Thus provided that the ratio of estimated likelihoods approximates the true ratio "reasonably well", i.e. in some sense

$$\frac{\hat{\mathcal{L}}(\boldsymbol{q}^{(\star)})}{\hat{\mathcal{L}}(\boldsymbol{q}^{(t)})} \approx \frac{\mathcal{L}(\boldsymbol{q}^{(\star)})}{\mathcal{L}(\boldsymbol{q}^{(t)})}, \tag{6.15}$$

an "appropriate" decision is being made. High posterior fidelity is ensured on condition that "appropriate" decisions are being frequently made over the course of the Markov process, i.e. detailed balance is maintained in some average sense. That this is indeed the case depends on a complex interplay between the quality of the estimation $\hat{\mathcal{L}}$ of $\mathcal{L}$, the true posterior π_1 and the proposal distribution P.

Similar arguments have been invoked in connection with Bayesian inversion of MC forward models, where algorithms have been designed that make use of forward model evaluations at multiple MC resolutions [47]. Moreover approximate MH corrections have been proposed in the context of big data, where evaluating the likelihood for partial chunks of observations from a large dataset trades off the bias and the variance of MCMC

posterior sampling [48, 49]. In order to provide a likelihood simulator that does not exhibit variations between consecutive evaluations for the same arguments, it was proposed to employ a common random number generator seed [50]. Once the seed is chosen, likelihood evaluations become effectively deterministic. The pseudo-marginal approach to MCMC [51] provides a strong theoretical result regarding posterior fidelity. It is shown that for unbiased likelihood estimators the exact posterior can be sampled. Though this is accompanied by a slowdown in the mixing properties of the Markov chain. With that said one may suppose that it is preferable to minimize the bias in Eq. (6.12) instead of the total mean squared error.

6.5 The NASA Langley multidisciplinary UQ challenge

Now we will interpret the uncertainty characterization subproblem A of the NASA Langley UQ challenge [1, 2] in Bayesian terms. For that purpose we will compose an appropriate Bayesian multilevel model and formulate the main objective, i.e. the reduction of epistemic uncertainties, as Bayesian calibration of this multilevel model. Inputs $(p_1, p_2, p_3, p_4, p_5)$ of the forward model $\mathcal{M} \equiv h_1$ are subject to uncertainty. Model inputs $\boldsymbol{\zeta} \equiv p_3$, that are subject to aleatory variability, constitute the category I parameters. There is epistemic uncertainty about the true value of the category II model parameter $\boldsymbol{m} \equiv p_2$. Category III subsumes those parameters $\boldsymbol{x} \equiv (p_1, p_4, p_5)$ that are subject to a mixed-type uncertainty. Generally, population distributions of experiment-specific variables are provided by the organizers of the challenge problem, whereas prior marginals of the QoI are uninformative interpretations of the epistemic intervals given.

6.5.1 Category I: Aleatory uncertainty

For experiments $i = 1, \dots, n$ category I model inputs $\boldsymbol{\zeta} \equiv p_3 \in [0, 1]$ take on experiment-specific realizations $p_{3,i}$. The population distribution is a uniform distribution $\mathcal{U}(p_{3,i}|a_3, b_3)$ determined by perfectly known hyperparameters $\boldsymbol{\theta_Z} \equiv \boldsymbol{\theta}_3 = (a_3, b_3)$ with $(a_3, b_3) = (0, 1)$. We write this as follows

$$(P_{3,i}|\boldsymbol{\theta}_3) \sim f_3(p_{3,i}|\boldsymbol{\theta}_3) = \mathcal{U}(p_{3,i}|0, 1). \tag{6.16}$$

It corresponds to a prescribed aleatory variability or structural uncertainty that is irreducible in the sense that by analyzing available data $\tilde{y}_i$ for $i = 1, \dots, n$ "past" realizations $p_{3,i}$ could be inferred in principle, whereas the knowledge about "future" realizations $p_{3,i'}$ with $i' > n$ cannot be improved.

6.5.2 Category II: Epistemic uncertainty

Category II model inputs are physically fixed yet unknown model parameters $\boldsymbol{m} \equiv p_2 \in [0, 1]$. A given epistemic interval $\Delta = [0, 1]$ is known to contain the true value of p_2 prior to any data analysis. We translate this available information into a flat and uniform Bayesian prior probability density

$$P_2 \sim \pi_2(p_2) = \mathcal{U}(p_2|0, 1). \tag{6.17}$$

It represents an a priori degree of plausibility of the true value p_2 and it is reducible in the sense that Bayesian updating provides an a posteriori degree of evidence. The quantification of parametric Bayesian priors is a controversial business. Priors go beyond bare interval-like statements by assigning a relative probability structure over the set of admissible values. Thus more generally any prior distribution with nonzero support over the priorly admissible set Δ that vanishes elsewhere could be considered appropriate.

6.5.3 Category III: Mixed uncertainty

Category III comprises those model inputs $\boldsymbol{x} \equiv (p_1, p_4, p_5)$ that are subject to aleatory variability across experiments $i = 1, \dots, n$. The natural variability is parametrized by hyperparameters $\boldsymbol{\theta_X} \equiv (\boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ that are epistemically uncertain themselves. This is a mixed-type uncertainty model that in less Bayesian contexts is sometimes referred to as imprecise probability or distributional p-box [52, 53].

6.5.3.1 Unimodal beta

Model inputs $p_1 \in [0, 1]$ are distributed according to a unimodal beta distribution. Beta distributions $\text{Beta}(p_{1,i}|\alpha_1, \beta_1)$ are commonly parametrized by shape hyperparameters $\alpha_1, \beta_1 > 0$. Instead we herein parametrize the beta distribution $\text{Beta}(p_{1,i}|\mu_1, \sigma_1^2)$ by its mean μ_1 and variance σ_1^2. Thus with unknown hyperparameters $\boldsymbol{\theta}_1 \equiv (\mu_1, \sigma_1^2)$ experiment-specific realizations $p_{1,i}$ are drawn from the population distribution

$$(P_{1,i}|\boldsymbol{\theta}_1) \sim f_1(p_{1,i}|\boldsymbol{\theta}_1) = \text{Beta}(p_{1,i}|\mu_1, \sigma_1^2). \tag{6.18}$$

To begin with, given the shape parameters (α_1, β_1), the expected value $\mu_1 = \mathbb{E}[p_1]$ and the variance $\sigma_1^2 = \mathrm{Var}[p_1]$ of the density function $\mathrm{Beta}(p_{1,i}|\mu_1, \sigma_1^2)$ are given as

$$\mu_1 = \frac{\alpha_1}{\alpha_1 + \beta_1}, \quad \sigma_1^2 = \frac{\alpha_1 \beta_1}{(\alpha_1 + \beta_1)^2(\alpha_1 + \beta_1 + 1)}. \tag{6.19}$$

Conversely, given the statistical moments (μ_1, σ_1^2), the shape parameters (α_1, β_1) of the density function $\mathrm{Beta}(p_{1,i}|\alpha_1, \beta_1)$ can be obtained by

$$\alpha_1 = \left(\frac{\sigma_1^2 + \mu_1^2 - \mu_1}{\sigma_1^2}\right)(-\mu_1), \quad \beta_1 = \left(\frac{\sigma_1^2 + \mu_1^2 - \mu_1}{\sigma_1^2}\right)(\mu_1 - 1). \tag{6.20}$$

The required unimodality, i.e. the fact that the distribution features a single mode within its support, translates into $\alpha_1, \beta_1 > 1$. Moreover the problem setup requires $3/5 \leq \mu_1 \leq 4/5$ and $1/50 \leq \sigma_1^2 \leq 1/25$. In order to adopt this epistemic uncertainty model for the hyperparameters $\boldsymbol{\theta}_1$ we state the uniform hyperprior

$$\begin{aligned} &\boldsymbol{\Theta}_1 \sim \pi_1(\boldsymbol{\theta}_1) = \mathcal{U}(\boldsymbol{\theta}_1|\mathcal{D}_{\boldsymbol{\theta}_1}), \text{ with} \\ &\mathcal{D}_{\boldsymbol{\theta}_1} = \left\{(\mu_1, \sigma_1^2) \in \mathbb{R}^2 | 3/5 \leq \mu_1 \leq 4/5,\ 1/50 \leq \sigma_1^2 \leq 1/25,\ \alpha_1 > 1,\ \beta_1 > 1\right\}. \end{aligned} \tag{6.21}$$

If $\lambda(\mathcal{D}_{\boldsymbol{\theta}_1})$ is the volume of the set $\mathcal{D}_{\boldsymbol{\theta}_1} \subset \mathbb{R}^2$, then the uniform density Eq. (6.21) is $1/\lambda(\mathcal{D}_{\boldsymbol{\theta}_1})$ on $\mathcal{D}_{\boldsymbol{\theta}_1}$ and zero elsewhere. In practice the normalization constant $\lambda(\mathcal{D}_{\boldsymbol{\theta}_1})$ is unknown, but since priors are flat and only ratios are compared in the MH correction Eq. (6.14), only the set membership of MCMC proposals has to be determined.

Consequently we can treat the prior Eq. (6.21) as $\pi_1(\boldsymbol{\theta}_1) = \pi(\mu_1)\,\pi(\sigma_1^2)$ with independent marginals $\pi(\mu_1) = \mathcal{U}(\mu_1|3/5, 4/5)$ and $\pi(\sigma_1^2) = \mathcal{U}(\sigma_1^2|1/50, 1/25)$ and reject MCMC proposals that do not respect $\alpha_1, \beta_1 > 1$ with the aid of Eq. (6.20). This practical prior choice is ambiguous in the sense that priors could be assumed for shape parameters (α_1, β_1), too. However, this could yield improper prior distributions. From an engineering point of view, we consider (μ_1, σ_1^2) statistically more "natural" than the shape parameters. In addition they underlie strong prior constraints which is advantageous to exploring the posterior by means of MCMC.

6.5.3.2 Correlated Gaussian

The model inputs $p_4, p_5 \in \mathbb{R}$ are modeled as possibly correlated Gaussian random variables. Across the experiments $i = 1, \ldots, n$ these model inputs take on different unknown realizations $(p_{4,i}, p_{5,i})$. This inherently aleatory variability is represented by the population distribution

$$\big((P_{4,i}, P_{5,i})|\boldsymbol{\theta}_{45}\big) \sim f_{45}\big((p_{4,i}, p_{5,i})|\boldsymbol{\theta}_{45}\big) = \mathcal{N}\big((p_{4,i}, p_{5,i})|\boldsymbol{\mu}_{45}, \boldsymbol{\Sigma}_{45}\big). \tag{6.22}$$

For $j = 4, 5$ the means $\mu_j = \mathbb{E}[p_j]$, variances $\sigma_j^2 = \mathrm{Var}[p_j]$ and the coefficient of correlation $\rho_{45} = \mathbb{E}[(p_4 - \mu_4)(p_5 - \mu_5)]$ constitute the hyperparameters $\boldsymbol{\theta}_{45} \equiv (\mu_4, \sigma_4^2, \mu_5, \sigma_5^2, \rho_{45})$. Those hyperparameters are unknown constants that determine the mean $\boldsymbol{\mu}_{45}$ and the covariance matrix $\boldsymbol{\Sigma}_{45}$ of the bivariate normal density by

$$\boldsymbol{\mu}_{45} = \begin{pmatrix} \mu_4 \\ \mu_5 \end{pmatrix}, \quad \boldsymbol{\Sigma}_{45} = \begin{pmatrix} \sigma_4^2 & \rho_{45}\,\sigma_4\,\sigma_5 \\ \rho_{45}\,\sigma_4\,\sigma_5 & \sigma_5^2 \end{pmatrix}. \tag{6.23}$$

Besides the natural bounds $|\rho_{45}| \leq 1$ it is requested that $-5 \leq \mu_j \leq 5$ and $1/400 \leq \sigma_j^2 \leq 4$. We translate these intervals into flat and independent marginals $\pi(\mu_j)$, $\pi(\sigma_j^2)$ and $\pi(\rho_{45})$ of the common hyperprior $\pi_{45}(\boldsymbol{\theta}_{45})$ by

$$\left.\begin{aligned} \pi(\mu_j) &= \mathcal{U}(\mu_j| -5, 5), \\ \pi(\sigma_j^2) &= \mathcal{U}(\sigma_j^2|1/400, 4), \\ \pi(\rho_{45}) &= \mathcal{U}(\rho_{45}| -1, 1), \end{aligned}\right\} \quad \boldsymbol{\Theta}_{45} \sim \pi_{45}(\boldsymbol{\theta}_{45}) = \left(\prod_{j=4}^{5} \pi(\mu_j)\,\pi(\sigma_j^2)\right)\pi(\rho_{45}). \tag{6.24}$$

Insofar as priors for spread hyperparameters could refer to standard deviations or variances alike, the ambiguity in quantifying parametric Bayesian priors becomes especially obvious for these type of hyperparameters.

6.5.4 Bayesian problem statement

The primary objective of the NASA UQ challenge subproblem A is the reduction of epistemic uncertainties of the true values of the forward model parameter p_2 and the hyperparameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ [1]. In order to accomplish that goal, the forward model, a set of data and prior knowledge is available. Preventing to reverse-engineer its mathematical character and numerical implementation, the forward model h_1 is distributed as a

protected MATLAB p-code file, i.e. a "blackbox" model. Available data $\langle \tilde{y}_i \rangle_{1 \le i \le 50}$ comprises $n = 50$ scalar observations $\tilde{y}_i = h_1(p_{1,i}, p_2, p_{3,i}, p_{4,i}, p_{5,i})$ which have been realized as forward model responses complying with the true uncertainty model of forward model inputs, i.e. the model parameter p_2 takes on its true value and $(\langle p_{1,i} \rangle, \langle p_{3,i} \rangle, \langle p_{4,i} \rangle, \langle p_{5,i} \rangle)$ have been randomly sampled from their true population distributions. Notwithstanding that the observations provided are "perfect", in general they might very well be subject to an additional model-measurement discrepancy, i.e. "imperfect" [2]. Data have been arranged into two distinct configurations of observations $\langle \tilde{y}_i \rangle_{1 \le i \le 25}$ and $\langle \tilde{y}_i \rangle_{26 \le i \le 50}$ whose separate and joint analysis is envisaged to indicate how the number n of processed samples impacts the significance of the final results.

The available prior knowledge has been translated into parametric and structural Bayesian prior distributions. We have pointed out that this formulation endows the problem with a subjectivist interpretation of probability and suffers from the ambiguity in the chosen parametric prior and its influence on the resulting posterior. The problem statement as well as the framework and the algorithms introduced so far grant ample scope of formulating and solving the problem as Bayesian inference of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ within a multilevel context. In the first place the Bayesian multilevel model Eq. (6.7), defined by parametric priors Eqs. (6.17), (6.21) and (6.24) and structural priors Eqs. (6.16), (6.18) and (6.22), establishes the natural framework for solving the original problem posed. For the sake of completeness the devised multilevel model is summarized as

$$\begin{aligned}
(\tilde{Y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3) &\sim f(\tilde{y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3), \\
P_2 &\sim \pi_2(p_2) = \mathcal{U}(p_2 | 0, 1), \\
(P_{1,i} | \boldsymbol{\theta}_1) &\sim f_1(p_{1,i} | \boldsymbol{\theta}_1) = \mathrm{Beta}(p_{1,i} | \mu_1, \sigma_1^2), \\
\big((P_{4,i}, P_{5,i}) | \boldsymbol{\theta}_{45}\big) &\sim f_{45}\big((p_{4,i}, p_{5,i}) | \boldsymbol{\theta}_{45}\big) = \mathcal{N}\big((p_{4,i}, p_{5,i}) | \boldsymbol{\mu}_{45}, \boldsymbol{\Sigma}_{45}\big), \\
\boldsymbol{\Theta}_1 &\sim \pi_1(\boldsymbol{\theta}_1) = \mathcal{U}(\boldsymbol{\theta}_1 | \mathcal{D}_{\boldsymbol{\theta}_1}), \\
\boldsymbol{\Theta}_{45} &\sim \pi_{45}(\boldsymbol{\theta}_{45}) = \pi(\mu_4)\, \pi(\sigma_4^2)\, \pi(\mu_5)\, \pi(\sigma_5^2)\, \pi(\rho_{45}), \\
(P_{3,i} | \boldsymbol{\theta}_3) &\sim f_3(p_{3,i} | \boldsymbol{\theta}_3) = \mathcal{U}(p_{3,i} | 0, 1).
\end{aligned} \tag{6.25}$$

The corresponding posterior distribution Eq. (6.9) of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ follows Bayesian data analysis of the given forward model responses $\langle \tilde{y}_i \rangle$, i.e. realizations of random variables $(\tilde{Y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$.

In the second place one could solve the inverse problem posed in the presence of additional measurement noise. To that end the Bayesian multilevel model Eq. (6.1) establishes the proper framework. Synthetic and noisy observations $y_i = \tilde{y}_i + \varepsilon_i$ could be obtained by perturbing the given model responses $\tilde{y}_i$ with residuals ε_i that are randomly sampled from prescribed distributions $f_{E_i}(\varepsilon_i) = \mathcal{N}(\varepsilon_i | 0, \sigma_i^2)$. Parameters of the residual model, i.e. the residual variances σ_i^2, could either be treated as knowns or as further unknowns. By analyzing "imperfect" data $\langle y_i \rangle$, i.e. realizations of random variables $(Y_i | p_{1,i}, p_2, p_{3,i}, p_{4,i}, p_{5,i})$, and treating latent variables $(\langle p_{1,i} \rangle, \langle p_{3,i} \rangle, \langle p_{4,i} \rangle, \langle p_{5,i} \rangle)$ as nuisance, inference of the QoI would be based on the posterior Eq. (6.4). A DAG of the Bayesian multilevel model corresponding to our challenge problem interpretation with "perfect" and "imperfect" data, respectively, is depicted in Fig. 6.2.

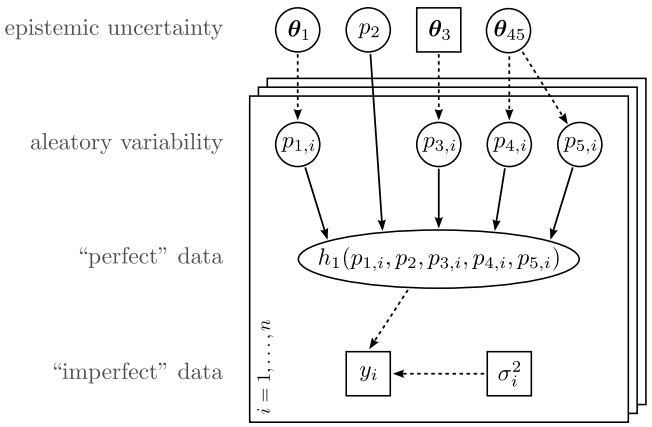


Figure 6.2: DAG of the NASA UQ challenge subproblem A. The hyperparameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_{45}$ and the forward model parameter p_2 located at the "highest" hierarchical level are the QoI. Realizations $(\langle p_{1,i} \rangle, \langle p_{3,i} \rangle, \langle p_{4,i} \rangle, \langle p_{5,i} \rangle)$ on the "intermediate" problem level are considered nuisance. "Perfect" $\tilde{y}_i = h_1(p_{1,i}, p_2, p_{3,i}, p_{4,i}, p_{5,i})$ or "imperfect" data $y_i = \tilde{y}_i + \varepsilon_i$ constitute the "lowest" model layer.

6.6 Bayesian data analysis

We will now apply the inferential machinery of multilevel calibration for solving the Bayesian interpretation of the uncertainty characterization subproblem A of the NASA Langley multidisciplinary UQ challenge. The problem will be solved in its original formulation involving "perfect" data. Motivated by findings from first preliminary problem analyses, posterior densities of the QoI will be computed by a suitable MH independence sampler. This sampler will be implemented in MATLAB and serially run on a modern CPU. Nevertheless we will discuss possible parallelization strategies. The total data $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$ and its subconfigurations $\langle \tilde{y}_i \rangle_{1 \leq i \leq 25}$ and $\langle \tilde{y}_i \rangle_{26 \leq i \leq 50}$ will be analyzed with the devised algorithm. Based on heuristic parameter tuning and plausibility checks we will assess the fidelity of the posterior. Promising a boost of posterior fidelity we will lastly devise a hybrid MCMC scheme which is based on data augmentation and both independence and random walk sampling.

6.6.1 Preliminary analyses

A basic understanding of an inverse problem under consideration allows to judge the performance of various potential MCMC schemes. This allows to design efficient algorithms and it is indispensable since it prevents from obtaining misleading results that are due to inappropriate samplers. In order to gain first insights into the present multilevel calibration problem, we perform a number of initial MCMC runs that were based on crude random walk Metropolis sampling. Thereby we could provisionally assess the principal nature of the posteriors Eqs. (6.3) and (6.9). Main findings from sampling the posterior Eq. (6.9) indicate that posterior marginals of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ may very well be multimodal or broad distributions that significantly overlap with the marginal priors.

Solving a joint problem in the presence of additional measurement noise provides further insight. Notwithstanding that this is actually a different problem, it will eventually prove valuable. Sampling the joint posterior Eq. (6.3) of the entirety of unknowns $(\langle p_{1,i} \rangle, p_2, \langle p_{3,i} \rangle, \langle p_{4,i} \rangle, \langle p_{5,i} \rangle, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ reveals further information about the unknowns, e.g. it occurs that experiment-specific unknowns $\langle p_{1,i} \rangle$ are identifiable and their posterior marginals feature single modes. This does not refer to the posteriors of $\langle p_{3,i} \rangle$ and $\langle p_{4,i} \rangle$ that are rather flat and the ones of $\langle p_{5,i} \rangle$ that are bimodal. Altogether those preliminary analyses have provided useful information that will eventually motivate the final MCMC samplers.

6.6.2 "Perfect" data analysis

For the calibration of the Bayesian multilevel model Eq. (6.7) we devise a blockwise independence MCMC sampler. Since the algorithm is based on MCMC, MC and KS techniques, hereinafter it will be referred to as MC^{3}KS. QoI are grouped in blocks (p_2), (μ_1, σ_1^2), $(\mu_4, \sigma_4^2, \rho_{45})$ and (μ_5, σ_5^2) that are consecutively updated by sampling blockwise candidates from the corresponding prior distributions. In many cases independence sampling from the priors is inefficient due to a negligible overlap between the prior and the posterior distributions and the resulting low acceptance rates. However, on account of the multimodality of the posteriors and their overlap with the priors, that were indicated by first analyses, independence sampling promises rapid mixing for the problem at hand.

Moreover in the context of Eq. (6.15) we suppose that wide jumps in the parameter space, that are induced by independence sampling on average, are beneficial in terms of posterior fidelity. For wide jumps the difference of the likelihood at the current and the candidate state of the Markov chain tends to be larger than for small jumps. Following previous discussions, to some extent this alleviates the error statistics of repeated likelihood estimations for the same state. Another advantage of the devised MCMC scheme over random walk sampling is that apart from its blockwise updating structure, it does not require extensive fine-tuning of the proposal distribution. Updating in blocks intents to minimize the number of calls to the likelihood Eq. (6.13) that are necessary for each block in each MCMC iteration, while maintaining high acceptance rates that are favorable for independence sampling. With the help of Eq. (6.20) the constraints $\alpha_1, \beta_1 > 1$ are enforced by rejecting nonconforming proposals in the block (μ_1, σ_1^2). The MC^{3}KS sampler is initialized by setting parameters in the middle of their admissible intervals. Due to rapid mixing the initialization is not of crucial importance for the employed sampling scheme. Generally speaking we expect that forward model parameters and mean hyperparameters are easier to identify than spread or even correlation hyperparameters.

6.6.3 Likelihood estimation and posterior fidelity

For the estimation Eq. (6.13) of the transformed likelihood Eq. (6.8) we choose kernel functions $\mathcal{K}$ of Gaussian type. In order to achieve a convenient trade-off between the conflicting endeavors fidelity of the posterior and ease of its computation, the number of samples K and the bandwidth h have to be set. In practice computational

resource limitations restrict the total number of affordable forward model runs, hence we approach parameter tuning from the situation of a given K.

Owing to the absence of a rigorous means to define a corresponding and "optimal" bandwidth h, we study the posteriors obtained for fixed $K = 10^4$ and decreasing h in a cascade of runs. We observe an initial shrinkage of the posterior, i.e. evolving from the flat prior it takes on definite shape, and an eventual collapse, i.e. the posterior flattens out again and looses its structure. The initial shrinkage is associated with significant changes of the posterior shape, the eventual breakdown is QoI-dependent, and in between the posterior is relatively stable with respect to h. We remark that this behavior is consistent with Eqs. (6.14) and (6.15). Significant oversmoothing the target density Eq. (6.10), i.e. a strongly biased estimator Eq. (6.13), can falsely assign posterior mass to QoI-values that do not well-explain or even contradict the data. Considerable undersmoothing of the target density, i.e. a high variance of the estimator Eq. (6.13), can cause "arbitrary" acceptances in the MH correction. We speculate that in between those extremes, the more stable the posterior is with respect to small changes in h, the more confident we can be to have revealed the true posterior. Beyond that we presume that a high degree of distinctiveness of the posterior with respect to the prior indicates high posterior fidelity. The converse statement does not hold, though.

In addition to those heuristics we perform a plausibility check as follows. During preliminary analysis we have solved the UQ challenge problem in the presence of additional measurement noise ε_i with $E_i \sim \mathcal{N}(\varepsilon_i|0, \sigma_i^2)$, i.e. we sampled a joint posterior of the form Eq. (6.3). If the corresponding noise-level σ_i^2 tends to zero the results of analyzing "imperfect" data should approach the ones of analyzing "perfect" data. Indeed we find that for low levels of noise $\sigma_i^2 \gtrsim 0$ the joint problem solution resembles our final results for analyzing "perfect" data. While the posterior Eq. (6.9) can only be approximately explored with the dubious aid of statistical likelihood estimations Eq. (6.13), the joint posterior Eq. (6.3) can be sampled exactly. Thus we have found that our approximate solution to the actual problem reminds of an exact solution to an only slightly different problem. For "well-behaved" problems we take this observation as an indication of an acceptable degree of posterior fidelity.

Following this discussion $K = 10^4$ and $h = 0.002$ constitutes our final parameter setup. The principle of estimating the density Eq. (6.10) and the transformed likelihood Eq. (6.8) is visualized in Fig. 6.3. Samples of $K = 10^4$ and $K = 10^7$ forward model responses are simulated for two different (hyper)parameter values $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{high}}$ and $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{low}}$. As judged from our final results, these are (hyper)parameter values of high and low degree of posterior evidence, respectively. For the smaller sample with $K = 10^4$ estimates of the sought densities $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ are shown. For reference purposes a histogram of the larger sample with $K = 10^7$ is shown. It can be seen that response densities $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ for $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{high}}$ and $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{low}}$ significantly overlap. This is a problem characteristic that complicates the statistical identification of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$.

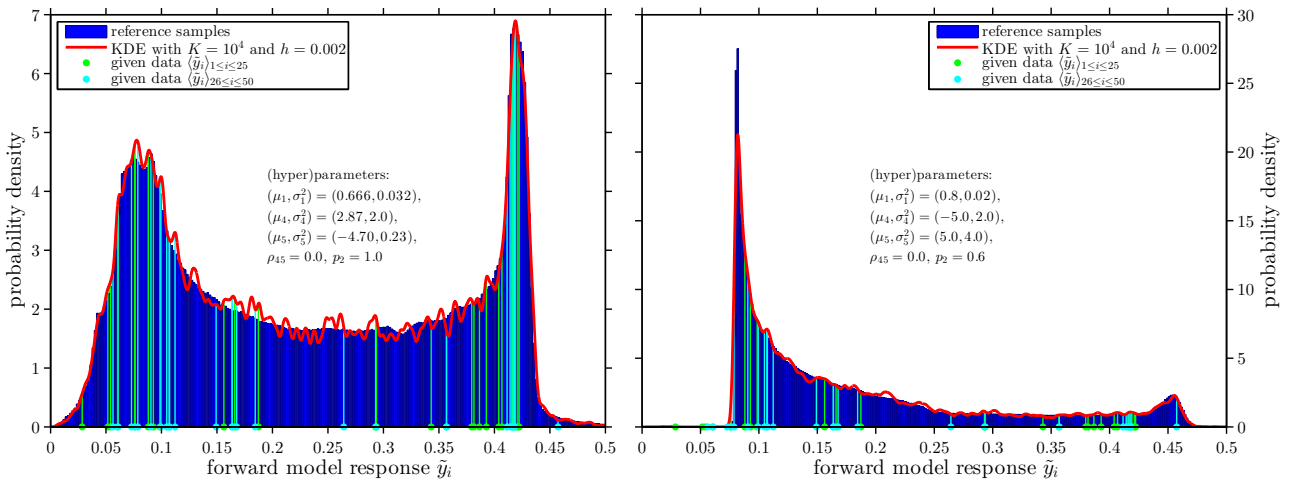


Figure 6.3: Estimation of $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$. Evaluating the transformed likelihood Eq. (6.8) for MC^{3}KS is based on the forward model response density Eq. (6.10). For two different values of the (hyper)parameters $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ a KDE of $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ with $K = 10^4$ and $h = 0.002$ is shown. Histograms with $K = 10^7$ forward model responses are shown as a reference.

It can also be seen that the employed bandwidth $h = 0.002$ amounts to a slight undersmoothing of the target density, i.e. a bias-variance trade-off favoring lower bias yet acceptable variance. This is advantageous because it allows to capture local small-scale features of the target density, e.g. sharp peaks and edges, in the posterior.

We remark that in the context of pseudo-marginal MCMC [51] this observation supports the speculation that it is preferable to minimize the bias in Eq. (6.12). Since the target density significantly differs from a normal distribution, automatic bandwidth selection cannot be based on the normal reference rule. The resulting oversmoothing of the target density, i.e. a significantly biased KDE, would veil its important characteristics. Finally the (hyper)parameter values $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{high}}$ can be seen to lead to a response density that explains the data sample $\langle \tilde{y}_i \rangle$ reasonably well.

6.6.4 Final results

First of all we jointly analyze the total data $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$. For $N = 10^5$ iterations of the MC^{3}KS algorithm the total program runtime amounts to $t \approx 30\,\text{h}$ on a single core. Blockwise acceptance rates were found to be ca. 20 % for (p_2), 40 % for (μ_1, σ_1^2), 60 % for $(\mu_4, \sigma_4^2, \rho_{45})$ and 10 % for (μ_5, σ_5^2). With Eq. (6.19) a number of 10327 blockwise proposals (μ_1, σ_1^2) had been rejected because of violating the prior requirement $\alpha_1, \beta_1 > 1$. Marginal posterior densities of the QoI are shown in Figs. 6.4 to 6.7. Based on appropriate boundary correction methods, the densities shown have been obtained by kernel smoothing of the MCMC posterior samples. Following precursory discussions we attribute an acceptable degree of fidelity to the posteriors obtained. We are confident that we have revealed a "good" approximation of the true posteriors, regardless of whether some of them are flat and only weakly informative.

We also analyze the data subconfigurations $\langle \tilde{y}_i \rangle_{1 \leq i \leq 25}$ and $\langle \tilde{y}_i \rangle_{26 \leq i \leq 50}$ separately. The posterior densities produced by separate analyses may differ considerably. With respect to the posteriors yielded by analyzing $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$, the two data subconfigurations are representative to a different degree. Those findings indicate that $n = 25$ is a comparably low number of observations while $n = 50$ is moderately satisfying for the Bayesian calibration of mean hyperparameters and the forward model parameter. Properly identifying the variance and correlation hyperparameters would require a higher number of observations. This is hardly surprising regarding the complex uncertainty setup, the number of unknowns, the unknown character of the forward model, and the inverse nature of the calibration problem.

At this point it is important to mention that multilevel model calibration shares and combines aspects of classical inverse problems, i.e. the inference of an unknown constant forward model parameter, and direct sample statistics, i.e. fitting a parametric distribution to a random data sample. Thereby Bayesian multilevel model calibration also inherits the usual difficulties inherent in inversion and distribution fitting.

The marginal posteriors of μ_1 and σ_1^2 are shown in Fig. 6.4. In comparison to the prior, the posterior of μ_1 shows a pronounced structure. Separately analyzing $\langle \tilde{y}_i \rangle_{1 \leq i \leq 25}$ and $\langle \tilde{y}_i \rangle_{26 \leq i \leq 50}$ gives rise to two different posterior modes. Those are suggested by the corresponding experiment-specific realizations $\langle p_{1,i} \rangle_{1 \leq i \leq 25}$ and $\langle p_{1,i} \rangle_{26 \leq i \leq 50}$. A joint analysis of $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$ leads to a mode that lies in between the two abovementioned ones. The posterior marginal of the spread hyperparameter σ_1^2 is comparably structureless and therefore less informative. In Fig. 6.5 the posterior marginals of p_2 and ρ_{45} are depicted. Due to the fact that data subconfigurations $\langle \tilde{y}_i \rangle_{1 \leq i \leq 25}$ and $\langle \tilde{y}_i \rangle_{26 \leq i \leq 50}$ are to a different degree informative about further unknowns, e.g. about (μ_1, σ_2^2) which was discussed above, the posteriors obtained for the constant model parameter p_2 may deviate as well. The analysis of $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$ by MC^{3}KS reveals two clear and separated posterior modes in the posterior of the forward model parameter p_2, whereas the posterior of the correlation hyperparameter ρ_{45} is flat. This is according to our previous expectations. Posteriors of μ_4 and σ_4^2 are given in Fig. 6.6. That they are comparably flat and uninformative prevents from drawing clear inferential conclusions. This statement does not hold for the posteriors of μ_5 and σ_5^2 that can be seen in Fig. 6.7. While the former features a bimodal structure and drops to zero for higher values of μ_5, the latter is unimodal and reaches a nonzero "plateau" for higher values of σ_5^2. Since this region accumulates considerable posterior probability mass, one cannot exclude those values of σ_5^2. Apart from the results of MC^{3}KS the figures Figs. 6.4 to 6.7 also contain the results of analyzing $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$ by MC^{3}DA, i.e. an alternative algorithm which is based on data augmentation. This technique will be further detailed in Section 6.7.

Note that Bayesian probabilities feature a richer structure than mere epistemic intervals. Conforming with a subjective Bayesian paradigm, probabilities are identified as relative degrees of belief or plausibility. Thus multivariate probability distributions, that may contain complex dependency structures and that are not entirely defined by their marginals only, have to be interpreted accordingly. The marginal densities shown hide this possibly existing posterior correlations. We provide a selection of two-dimensional posterior projections in Figs. 6.8 and 6.9.

Parameters that were assumed to be statistically independent a priori, e.g. the parameter p_2, the hyperparameters $\boldsymbol{\theta}_1$ and the hyperparameters $\boldsymbol{\theta}_{45}$, can be statistically dependent a posteriori. Small negative correlations in the posteriors of (μ_1, σ_1^2) and (μ_1, μ_4) shown in Figs. 6.8 and 6.9, with linear Pearson coefficients of correlation $r_{\mu_1, \sigma_1^2} = -0.08$ and $r_{\mu_1, \mu_4} = -0.22$ were discovered, respectively. In order to provide final results of interval-like

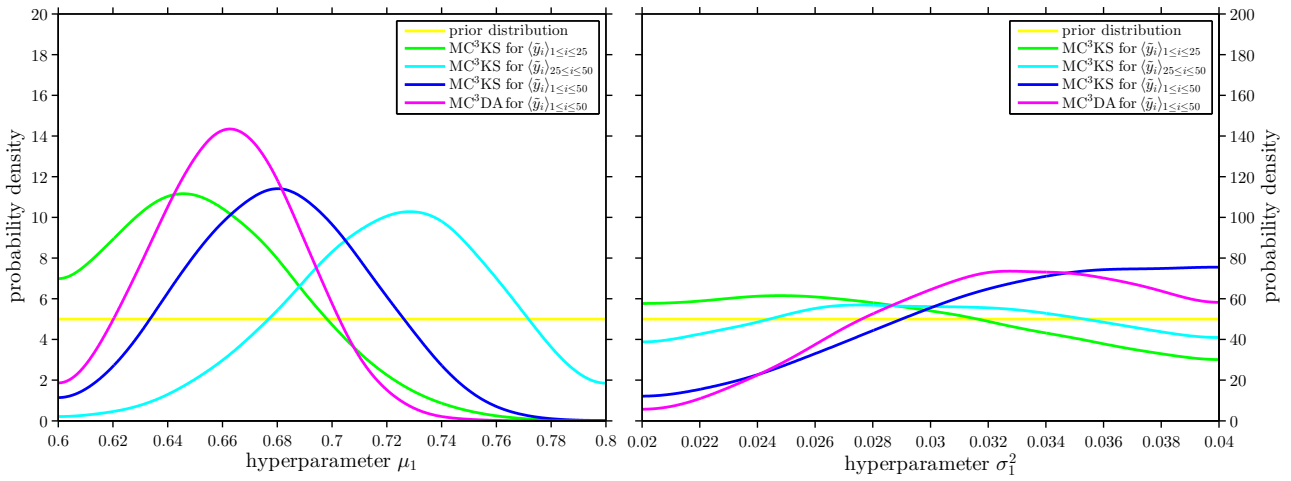


Figure 6.4: Posterior marginals of μ_1 and σ_1^2. The posterior of μ_1 features a clear structure as compared to the prior. Separate analyses of $\langle \tilde{y}_i \rangle_{1 \le i \le 25}$ and $\langle \tilde{y}_i \rangle_{26 \le i \le 50}$ lead to different posterior modes, whereas jointly analyzing $\langle \tilde{y}_i \rangle_{1 \le i \le 50}$ leads to a mode in between the two aforementioned ones. The posterior marginal of σ_1^2 is seen to be less informative.

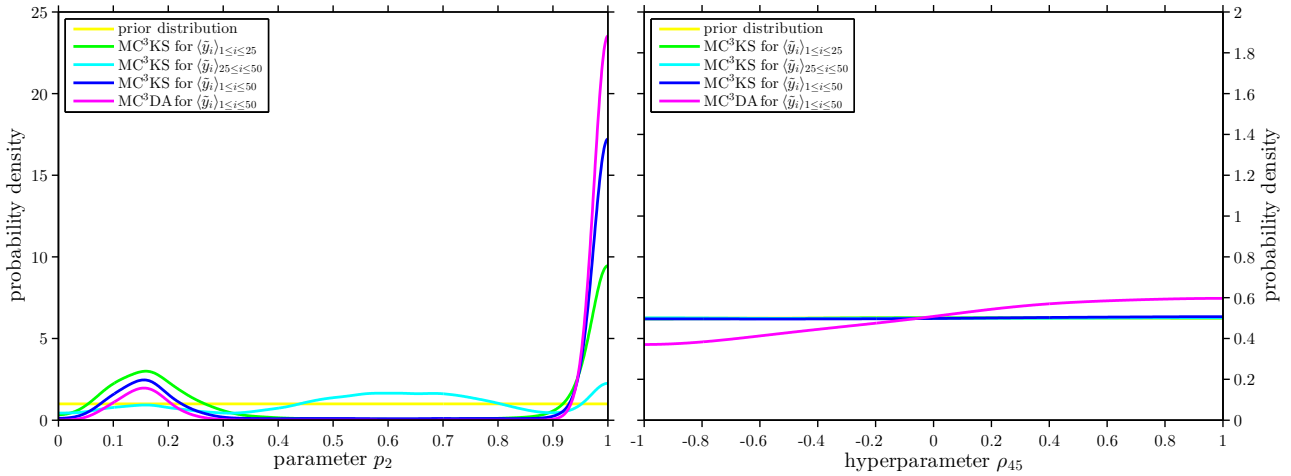


Figure 6.5: Posterior marginals of p_2 and ρ_{45}. Analyzing $\langle \tilde{y}_i \rangle_{1 \le i \le 50}$ by MC^{3}KS reveals two separated posterior modes in the posterior of p_2. As expected the posterior of the correlation hyperparameter ρ_{45} is flat and uninformative. Slightly more pronounced posterior structures are discovered by MC^{3}DA.

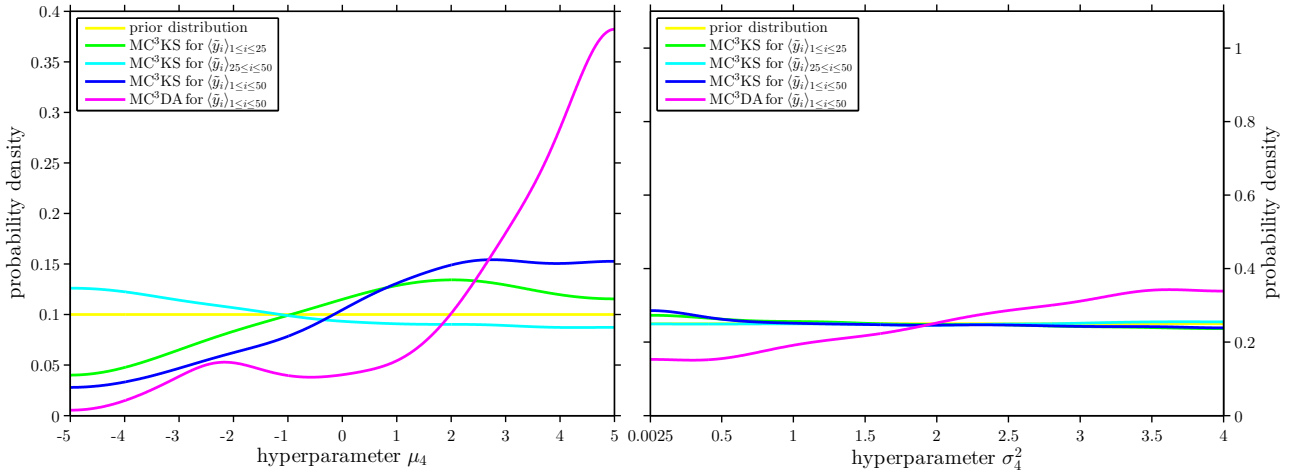


Figure 6.6: Posterior marginals of μ_4 and σ_4^2. Both the posterior marginals of μ_4 and σ_4^2 that were sampled by MC^{3}KS are rather flat and uninformative. On the contrary, the posterior of μ_4 explored by MC^{3}DA features more definite structure.

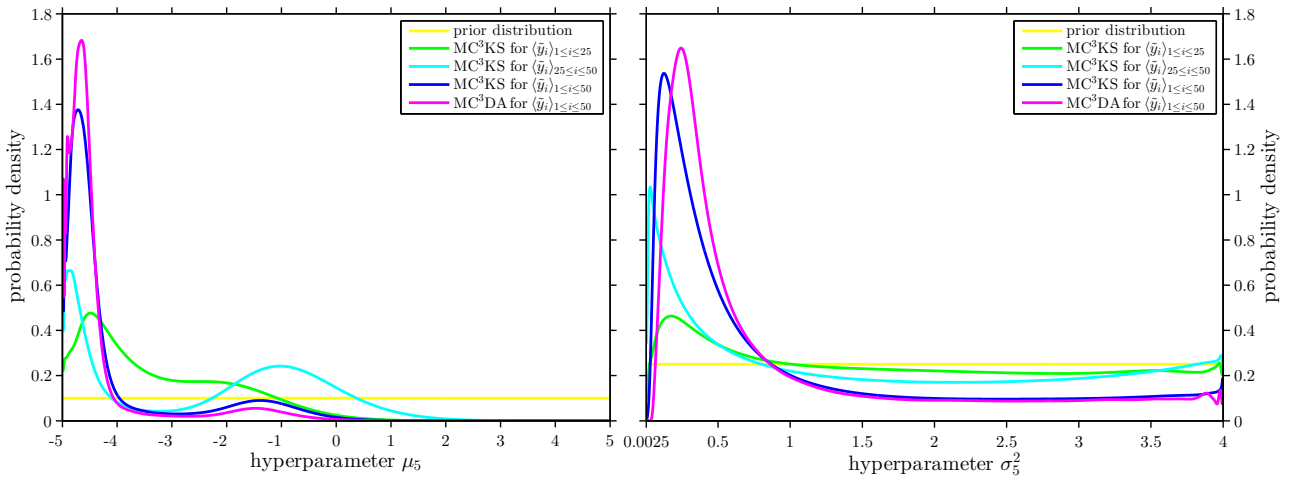


Figure 6.7: Posterior marginals of μ_5 and σ_5^2. The posterior marginals of μ_5 and σ_5^2 feature a distinctive structure as compared to the priors. The posterior of μ_5 is multimodal whereas the one of σ_5^2 is unimodal. With respect to the posteriors sampled by MC^{3}KS, the ones that are due to MC^{3}DA are slightly more evolved in structure.

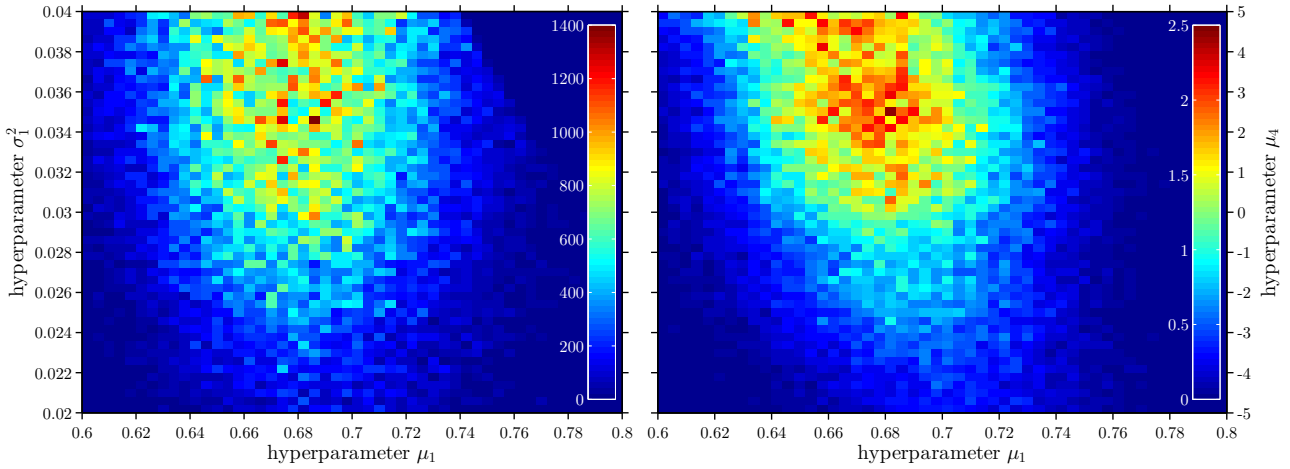


Figure 6.8: 2D posteriors of (μ_1, σ_1^2) and (μ_1, μ_4). Two-dimensional posterior projections for analyzing $\langle\tilde{y}_i\rangle_{1\le i\le 50}$ by MC^{3}KS are shown for (μ_1, σ_1^2) and (μ_1, μ_4). A small dependency structure in the posteriors can be seen. Linear Pearson coefficients of correlation are found as $r_{\mu_1,\sigma_1^2} = -0.08$ and $r_{\mu_1,\mu_4} = -0.22$.

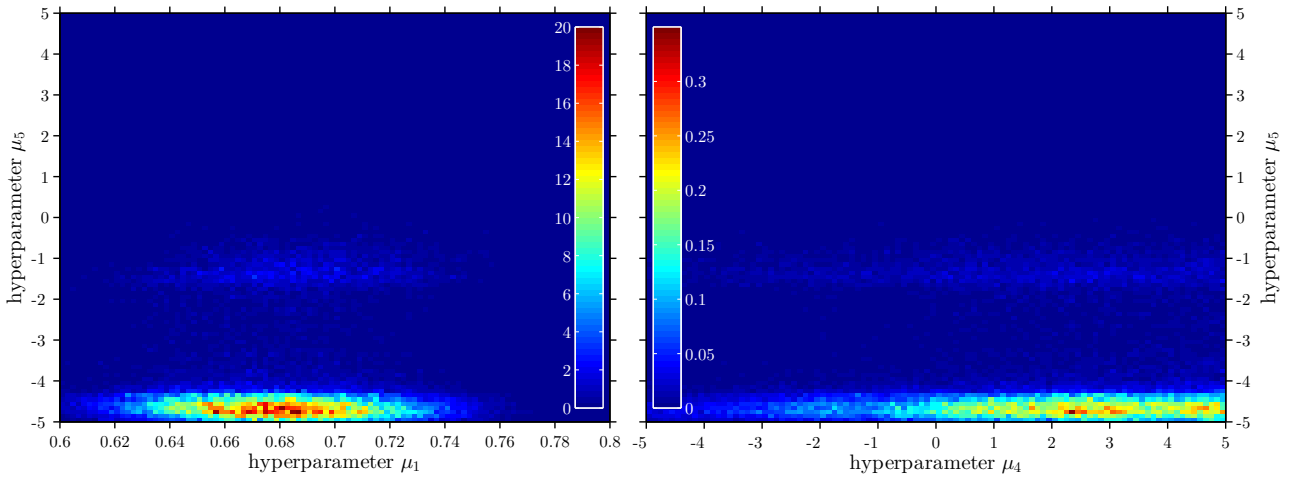


Figure 6.9: 2D posteriors of (μ_1, μ_5) and (μ_4, μ_5). For (μ_1, μ_5) and (μ_4, μ_5) two-dimensional posterior projections are shown that are due to analyzing $\langle\tilde{y}_i\rangle_{1\le i\le 50}$ with MC^{3}KS. The corresponding posteriors do not feature any marked dependency structure.

character one could define suitable Bayesian credible intervals or sets that accumulate a certain proportion, e.g. 95 %, of the total posterior mass. However, the definition of such intervals is ambiguous and would still bear the probabilistic interpretation, therefore we refrain from defining Bayesian credible intervals.

6.6.5 First conclusion

With the proposed MC^{3}KS algorithm the Bayesian formulation of the challenge problem could be solved. The numerical efficiency of the MCMC sampling scheme, that was based on independently sampling from the priors, could be easily increased. Obtained posteriors could be approximated by suitable distributions that are easy to sample. Utilizing these as proposal distributions would lead to higher acceptance rates and better mixing properties. Most Bayesian computations can only be parallelized by running several Markov chains simultaneously. An obvious parallelization strategy for the devised algorithm is to parallelize the estimation of the transformed likelihood on the level of forward model runs. This also suggests the possibility of studying the posterior for significantly larger K and smaller h. Moreover different classes of kernel functions $\mathcal{K}$, e.g. with bounded nonzero support, or more advanced KDE techniques, e.g. locally adaptive schemes or other bias reduction and correction methods, could be employed. The major shortcoming of the approach was the dependency of the final results on free algorithmic tuning parameters. Parameter tuning had to be based on heuristic criteria and plausibility checks and the fidelity of the final posteriors could only be provisionally assessed. In the following section we will therefore propose a complementary multilevel approach that aims at enhancing the level of posterior fidelity.

6.7 Partial data augmentation

As a potential improvement over the employed MC^{3}KS sampler we will devise a new hybrid MCMC sampling scheme. Since the scheme will be based on data augmentation (DA), henceforth it will be referred to as MC^{3}DA. Traditionally DA can be a powerful tool for enhancing the computational efficiency of MCMC posterior sampling [54–56]. Instead we will herein utilize DA as a means to reformulate the multilevel model calibration problem in such a complementary way, that it allows for more adequate likelihood estimations in view of Eq. (6.12). In turn this promises an enhancement of the posterior fidelity through Eq. (6.14). The approach will also allow for automatic kernel bandwidth selection based on a classical yet well-approved criterion, namely the normal reference rule [35]. This is appealing since it avoids the cumbersome procedure of tuning free algorithmic parameters of the KDE that was described in Section 6.6.3.

Rather than directly sampling the posterior of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$, one can sample the posterior of an augmented number of unknowns $(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ and obtain the posterior of the QoI by subsequently marginalizing over nuisance $\langle p_{1,i}\rangle$. Presuming that sampling from $\pi(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3)$ is "easier" to accomplish than straightforwardly sampling from $\pi(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3)$, a de facto improvement is achieved. The introduction of $\langle p_{1,i}\rangle$ as auxiliary variables is a partial form of data augmentation. As indicated by preliminary problem analyses, the forward model h_1 seems to be in such a strong way dependent on its input p_1, that the data $\langle\tilde{y}_i\rangle$ can be inverted for the unknown $\langle p_{1,i}\rangle$, under uncertainty of the remaining unknowns. Even though $\langle p_{1,i}\rangle$ are not QoI this provides additional insight into to inverse problem posed. Moreover the likelihood function corresponding to partial data augmentation can be estimated more adequately. Presumably, within a feasible computation time, the aforementioned facts will allow to sample $\pi(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3)$ with higher fidelity than sampling $\pi(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3)$. We will introduce the formalism of partial data augmentation below.

6.7.1 Augmented multilevel model

If the unknowns $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ of the Bayesian multilevel model Eq. (6.7) are augmented by experiment-specific realizations $\langle p_{1,i}\rangle$, then the collective of unknowns $(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ has to be explicitly taken into account. The associated Bayesian prior is given as

$$\pi\big(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}\big) = \left(\prod_{i=1}^{n} f_1(p_{1,i}|\boldsymbol{\theta}_1)\right) \pi_2(p_2)\, \pi_1(\boldsymbol{\theta}_1)\, \pi_{45}(\boldsymbol{\theta}_{45}). \tag{6.26}$$

This distribution comprises both parametric and structural prior knowledge. Given an appropriate probability model $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ of random variables $(\tilde{Y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$, the corresponding *augmented likelihood* follows as

$$\mathcal{L}\big(\langle\tilde{y}_i\rangle|\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3\big) = \prod_{i=1}^{n} f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3). \tag{6.27}$$

The definition of the augmented likelihood Eq. (6.27) rests upon a probability model $(\tilde{Y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3) \sim f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$. Analogous to the propagated uncertainty in Eq. (6.10), such a model is established through a transformed random variable $\mathcal{M}_{p_{1,i},p_2}(P_3, P_4, P_5)$ with a density function

$$f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3) = \int_0^1 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \delta\big(\tilde{y}_i - \mathcal{M}_{p_{1,i},p_2}(p_3, p_4, p_5)\big)\, f_3(p_3|\boldsymbol{\theta}_3)\, f_{45}\big((p_4, p_5)|\boldsymbol{\theta}_{45}\big)\, \mathrm{d}p_3\, \mathrm{d}p_4\, \mathrm{d}p_5. \tag{6.28}$$

Here δ denotes the Dirac delta function and $\mathcal{M}_{p_{1,i},p_2} \colon (p_3, p_4, p_5) \mapsto \mathcal{M}(p_{1,i}, p_2, p_3, p_4, p_5)$ formalizes the map that the forward model $\mathcal{M} \equiv h_1$ defines for fixed inputs $(p_{1,i}, p_2)$ and functional arguments (p_3, p_4, p_5). With the combined parametric and structural prior Eq. (6.26) and the augmented likelihood Eq. (6.27), the augmented posterior of the unknowns $(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ is according to Bayes' law proportional to

$$\pi\big(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3\big) \propto \mathcal{L}\big(\langle\tilde{y}_i\rangle|\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3\big)\, \pi\big(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}\big). \tag{6.29}$$

Since we are not interested in inferring experiment-specific realizations $\langle p_{1,i}\rangle$ per se, they are treated as nuisance. Similar to the marginalization Eq. (6.4) the posterior of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ is thus found by integrating the posterior Eq. (6.29) as follows

$$\pi\big(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3\big) = \int_0^1 \dots \int_0^1 \pi\big(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3\big)\, \mathrm{d}\langle p_{1,i}\rangle, \tag{6.30}$$

where $\mathrm{d}\langle p_{1,i}\rangle = \mathrm{d}p_{1,1} \dots \mathrm{d}p_{1,n}$ as before. Both of the distributions Eqs. (6.9) and (6.30) equivalently define the desired posterior $\pi(\boldsymbol{m}, \boldsymbol{\theta_X}|\langle\tilde{\boldsymbol{y}}_i\rangle, \boldsymbol{\theta_Z}) \equiv \pi(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3)$. While Eq. (6.9) straightforwardly conditions via $\pi(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle\tilde{y}_i\rangle, \boldsymbol{\theta}_3) \propto \mathcal{L}(\langle\tilde{y}_i\rangle|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)\, \pi(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$, Eqs. (6.29) and (6.30) provide a rearrangement where nuisance variables $\langle p_{1,i}\rangle$ are firstly factored out before they are eventually eliminated.

Even though this scheme of data augmentation leads to a higher-dimensional estimation problem, it may be computationally advantageous. In practice the marginal Eq. (6.30) can be computed by sampling the joint distribution Eq. (6.29) and simply discarding samples of $\langle p_{1,i}\rangle$, i.e. the multi-dimensional integral does not have to be calculated explicitly. Writing the compound distribution

$$f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3) = \int_0^1 f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)\, f(p_{1,i}|\boldsymbol{\theta}_1)\, \mathrm{d}p_{1,i} \tag{6.31}$$

establishes the connection between the transformed likelihood $\mathcal{L}(\langle\tilde{y}_i\rangle|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ of the form Eq. (6.8) and the augmented likelihood $\mathcal{L}(\langle\tilde{y}_i\rangle|\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ in Eq. (6.27). The relation Eq. (6.31) suggests that $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ could be a "simplified" version of the "complex" distribution $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ that was exemplarily shown in Fig. 6.3. Ideally it would be a unimodal distribution that resembles a Gaussian. Indeed in this case the augmented likelihood could be estimated more adequately than the transformed one. Consequently, the QoI-marginal of the induced limiting distribution of the MC^{3}DA scheme can be expected to be closer to the true posterior than the long-run distribution of the MC^{3}KS approach.

An augmented model can be analogously defined when latent variables other than $\langle p_{1,i}\rangle$ are introduced as auxiliary variables. The more variables are introduced, the smaller the variance of the target density similar to Eq. (6.28) is expected to be. In the case that all variables $(\langle p_{1,i}\rangle, \langle p_{3,i}\rangle, \langle p_{4,i}\rangle, \langle p_{5,i}\rangle)$ are jointly introduced, the conditional distribution of the data would even shrink to a Dirac delta $f(\tilde{y}_i|p_{1,i}, p_2, p_{3,i}, p_{4,i}, p_{5,i}) = \delta(\tilde{y}_i - \mathcal{M}_{p_{1,i},p_2,p_{3,i},p_{4,i},p_{5,i}})$. Note that for if "imperfect" data $y_i = \tilde{y}_i + \varepsilon_i$ would be involved, a proper distribution would still be defined as $f(y_i|p_{1,i}, p_2, p_{3,i}, p_{4,i}, p_{5,i}) = f_{E_i}(y_i - \mathcal{M}_{p_{1,i},p_2,p_{3,i},p_{4,i},p_{5,i}})$. In fact this exactly defines the distribution Eq. (6.1a) of the a joint problem in Eq. (6.1). The motivation for augmenting with $\langle p_{1,i}\rangle$ is the expectation that this provides a convenient trade-off between fidelity and feasibility, i.e. likelihood evaluations are optimally facilitated with respect to the implied increase in dimensionality.

6.7.2 Augmented likelihood estimation

In practical terms the augmented likelihood Eq. (6.27) can be estimated analogously to Eq. (6.13). To that end the response density $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is estimated for each $p_{1,i}$ and evaluated for the given responses $\tilde{y}_i$.

Hence a KDE-based estimate of the augmented likelihood as a function of $(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_{45})$ is given as

$$\hat{\mathcal{L}}_{\mathrm{DA}}\big(\langle \tilde{y}_i\rangle \,|\, \langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3\big) = \prod_{i=1}^{n} \left(\frac{1}{K} \sum_{k=1}^{K} \mathcal{K}_{h_i}\left(\tilde{y}_i - \tilde{y}_i^{(k)}\right)\right),$$
$$\text{with} \quad \begin{cases} p_{3,i}^{(k)} \sim f_3(p_{3,i}^{(k)}|\boldsymbol{\theta}_3), \\ (p_{4,i}^{(k)}, p_{5,i}^{(k)}) \sim f_{45}((p_{4,i}^{(k)}, p_{5,i}^{(k)})|\boldsymbol{\theta}_{45}), \\ \tilde{y}_i^{(k)} = \mathcal{M}_{p_{1,i},p_2}(p_{3,i}^{(k)}, p_{4,i}^{(k)}, p_{5,i}^{(k)}), \\ h_i = (4/3K)^{1/5}\,\hat{\sigma}_i. \end{cases} \tag{6.32}$$

For $k = 1, \ldots, K$ inputs $p_{3,i}^{(k)} \sim f_3(p_{3,i}^{(k)}|\boldsymbol{\theta}_3)$ and $(p_{4,i}^{(k)}, p_{5,i}^{(k)}) \sim f_{45}((p_{4,i}^{(k)}, p_{5,i}^{(k)})|\boldsymbol{\theta}_{45})$ are sampled from the corresponding population distributions, responses $\tilde{y}_i^{(k)} = \mathcal{M}_{p_{1,i},p_2}(p_{3,i}^{(k)}, p_{4,i}^{(k)}, p_{5,i}^{(k)})$ are computed accordingly. Furthermore $\hat{\sigma}_i$ denotes the standard deviation of the response samples $(\tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(K)})$. Note that in Eq. (6.32) the density $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is individually estimated for each $p_{1,i}$ with $i = 1, \ldots, n$. The number of samples for each of these estimations is set to $K = 10^3$ and selection of the bandwidths follows the normal reference rule $h_i = (4/3K)^{1/5}\,\hat{\sigma}_i$.

Let us compare the transformed densities Eqs. (6.10) and (6.28). The random variable $\mathcal{M}_{p_2}(P_1, P_3, P_4, P_5) \sim f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is conditioned on $\boldsymbol{\theta}_1$ and involves the uncertainty of $P_1 \sim f_1(p_1|\boldsymbol{\theta}_1)$. In contrast $\mathcal{M}_{p_{1,i},p_2}(P_3, P_4, P_5) \sim f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is conditioned on the realization $p_{1,i}$ and does not bear reference to $\boldsymbol{\theta}_1$. Hence $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is the broader and more complex density, whereas $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is simpler and easier to estimate. In turn, likelihood estimations for MC^{3}DA are less biased and have a smaller variance than for MC^{3}KS. Thus the approach promises a higher degree of posterior fidelity. In Fig. 6.10 the density $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is shown for values $(p_2, \boldsymbol{\theta}_{45})_{\mathrm{high}}$ and $(p_2, \boldsymbol{\theta}_{45})_{\mathrm{low}}$ that have been chosen as the same values already used in Fig. 6.3. Following our final results, the value $p_{1,i}$ has been exemplarily chosen as the posterior mean of $p_{1,i}$ with $i = 36$. When comparing Figs. 6.3 and 6.10 one can clearly see the essential difference between the densities $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ and $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$. The latter is distinctly simpler and clearly better resembling a Gaussian density. Moreover the densities for (hyper)parameter values of high and low posterior evidence only negligibly overlap. It can also be seen that the chosen values $p_{1,36}$ and $(p_2, \boldsymbol{\theta}_{45})_{\mathrm{high}}$ lead to a response density which is consistent with the observation $\tilde{y}_{36}$.

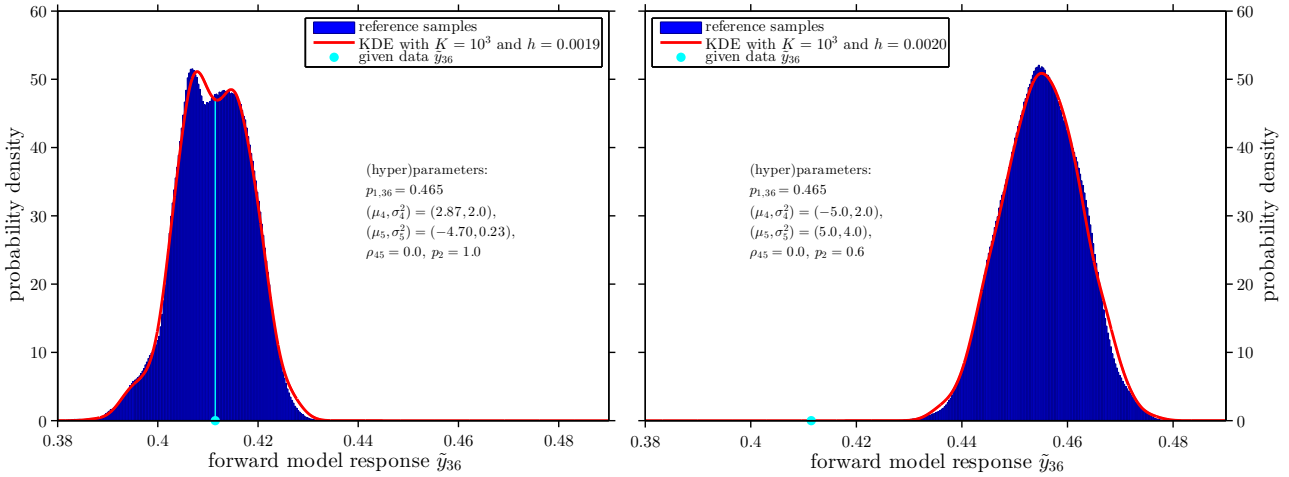


Figure 6.10: Estimation of $f(\tilde{y}_{36}|p_{1,36}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$. Evaluating the augmented likelihood Eq. (6.27) for MC^{3}DA is based on the response density Eq. (6.28). For $p_{1,36} = 0.465$ and two different values of the (hyper)parameters $(p_2, \boldsymbol{\theta}_{45})_{\mathrm{high}}$ and $(p_2, \boldsymbol{\theta}_{45})_{\mathrm{low}}$ a KDE of $f(\tilde{y}_{36}|p_{1,36}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ with $K = 10^3$ is shown. The bandwidths $h = 0.0019$ and $h = 0.0020$ were automatically selected according to the normal reference rule. Histograms with a larger number $K = 10^7$ are shown for reference purposes.

6.7.3 MCMC

The augmented posterior Eq. (6.29) is explored by means of a suitable MC^{3}DA sampler. Updating is done in blocks $\langle p_{1,i}\rangle$, (μ_1, σ_1^2), (p_2), (μ_4), (μ_5) and $(\sigma_4^2, \sigma_5^2, \rho_{45})$. Each $p_{1,i}$ in the block $\langle p_{1,i}\rangle$ is concurrently updated

with a random walk Metropolis sampler based on independent Gaussian proposals with standard deviation $\sigma_{p_{1,i}} = 0.01$. As before the remaining blocks are initialized in the middle of the corresponding epistemic intervals and updated with independent prior proposals. Acceptance rates amounted to ca. 10 % for $\langle p_{1,i} \rangle$, (p_2), and (μ_5), 15 % for (μ_4) and $(\sigma_4^2, \sigma_5^2, \rho_{45})$, and 30 % for (μ_1, σ_1^2). A number of 10219 proposals in the last-mentioned block were rejected due to violating $\alpha_1, \beta_1 > 1$. We start with preliminary MCMC runs with $K = 10^3$ and constant bandwidths $h_i = 0.02$ in order to identify the posterior modes of $\langle p_{1,i} \rangle$. Experiment-specific realizations $\langle p_{1,i} \rangle$ are initialized in the middle of their epistemic intervals and converge within ca. 1000 MCMC iterations. The initial convergence and final posterior of an experiment-specific realization $p_{1,i}$ with $i = 10$ are shown in Fig. 6.11. This shows that individual experiment-specific realizations $\langle p_{1,i} \rangle$ can indeed be inferred. The danger of the approach is that missing further posterior modes of $\langle p_{1,i} \rangle$ would alter the sampled posteriors of the remaining unknowns, above all the one of $\boldsymbol{\theta}_1$. Convergence checks have therefore been accomplished by initializing $\langle p_{1,i} \rangle$ within admissible regions of the parameter space that have not been visited in previous runs. Ultimately the chains converged to the same posterior modes which were found before. We conclude that the parameter space has been properly explored.

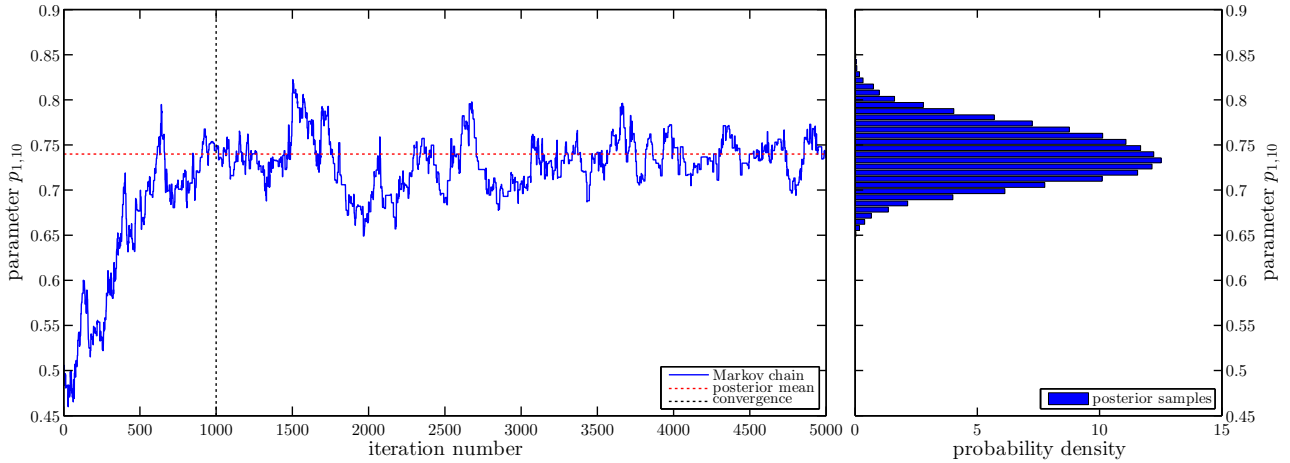


Figure 6.11: Convergence and identifiability of $p_{1,10}$. With $N = 10^5$ iterations of the MC^{3}DA algorithm the augmented posterior Eq. (6.29) is explored by analyzing $\langle \tilde{y}_i \rangle_{1 \le i \le 50}$. For a preliminary run with $K = 10^3$ and fixed $h_i = 0.02$, i.e. without automatic bandwidth selection, a trace plot of the converging Markov chain of an experiment-specific $p_{1,10}$ is shown. After convergence within ca. 1000 iterations the Markov chain samples the corresponding posterior around its mean.

We point out that although partial data augmentation has been motivated by considerations of posterior fidelity, it also gives additional insight into the inverse problems posed. Incidentally the posterior of experiment-specific realizations $\langle p_{1,i} \rangle$ is explored and its modes are identified. Thus we have gained knowledge about unknown problem quantities that are not primary QoI. Eventually we initialize the final sampler within the detected posterior modes of $\langle p_{1,i} \rangle$. With $K = 10^3$ and automatic selection of the bandwidths h_i we draw $N = 10^5$ posterior samples. Total execution time amounts to $t \approx 90$ h on a single core. The resulting posterior marginals of the QoI are added to Figs. 6.4 to 6.7. As compared to the results obtained by MC^{3}KS the posteriors found by MC^{3}DA have been slightly shrunk and evolved in structure. Resting upon the assumption that the posterior modes of $\langle p_{1,i} \rangle$ have been correctly identified, we take this as an indication of a gain in posterior fidelity. In Fig. 6.12 two-dimensional posteriors are shown for (μ_1, σ_1^2) and $(\mu_1, p_{1,i})$ with $i = 19$. The corresponding linear coefficients of correlation are found to be $r_{\mu_1, \sigma_1^2} = -0.25$ and $r_{\mu_1, p_{1,19}} = 0.19$. Generally we find small linear correlations $r_{\mu_1, p_{1,i}} \gtrsim 0$ between the mean hyperparameter μ_1 and experiment-specific realizations $p_{1,i}$ for nearly all $i = 1, \ldots, n$. This is plausible since higher values of μ_1 increase the plausibility of higher values of each $p_{1,i}$ and vice versa.

6.8 Conclusion and outlook

Addressing the uncertainty characterization subproblem of the NASA Langley multidisciplinary UQ challenge has turned out to be a challenging yet rewarding task. We began with formulating a generic Bayesian multilevel framework for managing different types of forward model input uncertainties in complex inverse problems. Incidentally this showed how the problem could be solved for "imperfect" data, e.g. in the presence of additional measurement noise, and how the entirety of problem unknowns, including those that are not of declared inferential

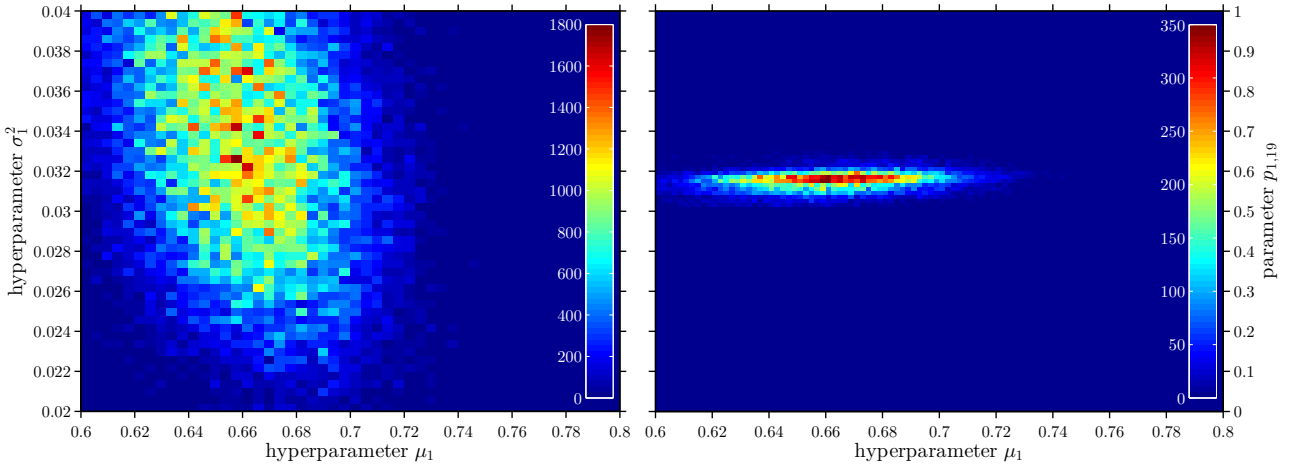


Figure 6.12: 2D posteriors of (μ_1, σ_1^2) and $(\mu_1, p_{1,19})$. Posterior projections that follow analyzing $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$ are shown for (μ_1, σ_1^2) and $(\mu_1, p_{1,19})$. While the former can be compared to the corresponding posterior marginal in Fig. 6.8 for MC^{3}KS, the latter is appertain to MC^{3}DA. Linear coefficients of correlation are found to be $r_{\mu_1,\sigma_1^2} = -0.25$ and $r_{\mu_1,p_{1,19}} = 0.19$.

interest, can be deduced. Although these were not the guiding questions, this is a future research direction in its own [57]. Bayesian structural prior modeling served as a foundation for devising a multilevel model in the zero-noise or "perfect" data limit, i.e. the data space has been endowed with a probability model that was based on uncertainty propagation. Ensuing from those general considerations we have interpreted and solved the challenge problem as Bayesian calibration of a suitably defined multilevel model. We thoroughly commented on the assumptions that the adopted approach rests upon as well as the interpretations it entails.

In turn the problem solution has given rise to new questions of theoretical and practical relevance alike. Posterior fidelity was discussed in the context of MCMC posterior exploration and online uncertainty propagation. First related thoughts were given and an in-depth consideration has been initiated. The starting point of the latter could be Eqs. (6.12) to (6.15). With the objective of improving the fidelity of the final results we demonstrated how one can exploit partial data augmentation. In addition to improving the estimation of the QoI, in principle this approach allows to infer such problem unknowns that inferential interest is not particularly focused on. That way partial data augmentation has provided further insight into the calibration problem posed. In sum we hope that these efforts prove to be a solid contribution to the NASA challenge problem in particular and to the theory and practice of Bayesian data analysis and uncertainty quantification in general. Future research work encompasses the design of more sophisticated methods to simulate the likelihood function and the rigorous assessment of the posterior fidelity.

References

[1] L. G. Crespo, S. P. Kenny, and D. P. Giesy. "The NASA Langley Multidisciplinary Uncertainty Quantification Challenge". In: *16th AIAA Non-Deterministic Approaches Conference (SciTech 2014)*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2014. DOI: 10.2514/6.2014-1347.

[2] L. G. Crespo, S. P. Kenny, and D. P. Giesy. *NASA LaRC UQ Challenge 2014: Challenge Problem FAQ*. February 19, 2013. URL: http://uqtools.larc.nasa.gov/nda-uq-challenge-problem-2014/faq/.

[3] J. V. Foster, K. Cunningham, C. M. Fremaux, G. H. Shah, E. C. Stewart, R. A. Rivers, J. E. Wilborn, and W. Gato. "Dynamics Modeling and Simulation of Large Transport Airplanes in Upset Conditions". In: *AIAA Guidance, Navigation, and Control Conference and Exhibit*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2005. DOI: 10.2514/6.2005-5933.

[4] T. L. Jordan and R. M. Bailey. "NASA Langley's AirSTAR Testbed: A Subscale Flight Test Capability for Flight Dynamics and Control System Experiments". In: *AIAA Guidance, Navigation and Control Conference and Exhibit*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2008. DOI: 10.2514/6.2008-6660.

[5] L. G. Crespo, M. Matsutani, and A. M. Annaswamy. "Design of an Adaptive Controller for a Remotely Operated Air Vehicle". In: *Journal of Guidance, Control, and Dynamics* 35.2 (2012), pp. 406–422. DOI: 10.2514/1.54779.

[6] M. H. Faber. "On the Treatment of Uncertainties and Probabilities in Engineering Decision Analysis". In: *Journal of Offshore Mechanics and Arctic Engineering* 127.3 (2005), pp. 243–248. DOI: 10.1115/1.1951776.

[7] A. Der Kiureghian and O. Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009), pp. 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

[8] R. Hadidi and N. Gucunski. "Probabilistic Approach to the Solution of Inverse Problems in Civil Engineering". In: *Journal of Computing in Civil Engineering* 22.6 (2008), pp. 338–347. DOI: 10.1061/(ASCE)0887-3801(2008)22:6(338).

[9] J. L. Beck. "Bayesian system identification based on probability logic". In: *Structural Control and Health Monitoring* 17.7 (2010), pp. 825–847. DOI: 10.1002/stc.424.

[10] A. M. Stuart. "Inverse problems: A Bayesian perspective". In: *Acta Numerica* 19 (2010), pp. 451–559. DOI: 10.1017/S0962492910000061.

[11] M. Davidian and D. M. Giltinan. "Nonlinear Models for Repeated Measurement Data: An Overview and Update". In: *Journal of Agricultural, Biological, and Environmental Statistics* 8.4 (2003), pp. 387–419. DOI: 10.1198/1085711032697.

[12] H. T. Banks, Z. R. Kenz, and W. C. Thompson. "A review of selected techniques in inverse problem nonparametric probability distribution estimation". In: *Journal of Inverse and Ill-posed Problems* 20.4 (2012), pp. 429–460. DOI: 10.1515/jip-2012-0037.

[13] E. de Rocquigny and S. Cambier. "Inverse probabilistic modelling of the sources of uncertainty: A non-parametric simulated-likelihood method with application to an industrial turbine vibration assessment". In: *Inverse Problems in Science and Engineering* 17.7 (2009), pp. 937–959. DOI: 10.1080/17415970902916987.

[14] G. Celeux, A. Grimaud, Y. Lefèbvre, and E. de Rocquigny. "Identifying intrinsic variability in multivariate systems through linearized inverse methods". In: *Inverse Problems in Science and Engineering* 18.3 (2010), pp. 401–415. DOI: 10.1080/17415971003624330.

[15] P. Barbillon, G. Celeux, A. Grimaud, Y. Lefèbvre, and E. de Rocquigny. "Nonlinear methods for inverse statistical problems". In: *Computational Statistics & Data Analysis* 55.1 (2011), pp. 132–142. DOI: 10.1016/j.csda.2010.05.030.

[16] J. Wakefield. "The Bayesian Analysis of Population Pharmacokinetic Models". In: *Journal of the American Statistical Association* 91.433 (1996), pp. 62–75. DOI: 10.1080/01621459.1996.10476664.

[17] D. J. Lunn. "Bayesian Analysis of Population Pharmacokinetic/Pharmacodynamic Models". In: *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Ed. by D. Husmeier, R. Dybowski, and S. Roberts. Advanced Information and Knowledge Processing. London: Springer, 2005, pp. 351–370. DOI: 10.1007/1-84628-119-9_11.

[18] M. J. Daniels. "A prior for the variance in hierarchical models". In: *Canadian Journal of Statistics* 27.3 (1999), pp. 567–578. DOI: 10.2307/3316112.

[19] A. Gelman. "Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper)". In: *Bayesian Analysis* 1.3 (2006), pp. 515–534. DOI: 10.1214/06-BA117A.

[20] J. C. Wakefield, A. F. M. Smith, A. Racine-Poon, and A. E. Gelfand. "Bayesian Analysis of Linear and Non-Linear Population Models by Using the Gibbs Sampler". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43.1 (1994), pp. 201–221. DOI: 10.2307/2986121.

[21] J. E. Bennett, A. Racine-Poon, and J. C. Wakefield. "MCMC for nonlinear hierarchical models". In: *Markov Chain Monte Carlo in Practice*. Ed. by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Interdisciplinary Statistics. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1995, pp. 339–357.

[22] J. B. Nagel and B. Sudret. "Probabilistic Inversion for Estimating the Variability of Material Properties: A Bayesian Multilevel Approach". In: *11th International Probabilistic Workshop (IPW11)*. Ed. by D. Novák and M. Vořechovský. Brno, Czech Republic: Litera, 2013, pp. 293–303. DOI: 10.3929/ethz-a-010034843.

[23] G. C. Ballesteros, P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos. "Bayesian Hierarchical Models for Uncertainty Quantification in Structural Dynamics". In: *2nd International Conference on Vulnerability and Risk Analysis and Management and 6th International Symposium on Uncertainty Modeling and Analysis (ICVRAM & ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 162, pp. 1615–1624. DOI: 10.1061/9780784413609.162.

[24] D. Draper, D. P. Gaver, P. K. Goel, J. B. Greenhouse, L. V. Hedges, C. N. Morris, and C. M. Waternaux. *Combining Information: Statistical Issues and Opportunities for Research*. Washington, D.C., USA: Panel on Statistical Issues et al., 1992.

[25] J. B. Nagel and B. Sudret. "A Bayesian Multilevel Approach to Optimally Estimate Material Properties". In: *2nd International Conference on Vulnerability and Risk Analysis and Management and 6th International Symposium on Uncertainty Modeling and Analysis (ICVRAM & ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 151, pp. 1504–1513. DOI: 10.1061/9780784413609.151.

[26] T. Koski and J. M. Noble. *Bayesian Networks: An Introduction*. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2009. DOI: 10.1002/9780470684023.

[27] U. B. Kjærulff and A. L. Madsen. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. 2nd ed. Information Science and Statistics 22. New York: Springer, 2013. DOI: 10.1007/978-0-387-74101-7.

[28] D. Basu. "On the Elimination of Nuisance Parameters". In: *Journal of the American Statistical Association* 72.358 (1977), pp. 355–366. DOI: 10.1080/01621459.1977.10481002.

[29] A. P. Dawid. "A Bayesian Look at Nuisance Parameters". In: *Trabajos de Estadística Y de Investigación Operativa* 31.1 (1980), pp. 167–203. DOI: 10.1007/BF02888351.

[30] J. O. Berger, B. Liseo, and R. L. Wolpert. "Integrated Likelihood Methods for Eliminating Nuisance Parameters". In: *Statistical Science* 14.1 (1999), pp. 1–28. DOI: 10.1214/ss/1009211804.

[31] T. A. Severini. "Integrated likelihood functions for non-Bayesian inference". In: *Biometrika* 94.3 (2007), pp. 529–542. DOI: 10.1093/biomet/asm040.

[32] J. B. Nagel and B. Sudret. "A Bayesian Multilevel Framework for Uncertainty Characterization and the NASA Langley Multidisciplinary UQ Challenge". In: *16th AIAA Non-Deterministic Approaches Conference (SciTech 2014)*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2014. DOI: 10.2514/6.2014-1502.

[33] P. Diaconis and D. Freedman. "On the Consistency of Bayes Estimates". In: *The Annals of Statistics* 14.1 (1986), pp. 1–26. DOI: 10.1214/aos/1176349830.

[34] S. J. Vollmer. "Posterior consistency for Bayesian inverse problems through stability and regression results". In: *Inverse Problems* 29.12, 125011 (2013), pp. 1–32. DOI: 10.1088/0266-5611/29/12/125011.

[35] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability 26. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1986.

[36] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Monographs on Statistics and Applied Probability 60. London, UK: Chapman & Hall/CRC, 1994.

[37] B. Sudret. "Uncertainty propagation and sensitivity analysis in mechanical models: Contributions to structural reliability and stochastic spectral methods". Habilitation à diriger des recherches. Clermont-Ferrand, France: Université Blaise Pascal, 2007.

[38] B. Sudret, F. Perrin, and M. Pendola. "Use of polynomial chaos expansions in stochastic inverse problems". In: *4th International ASRANet Colloquium*. Glasgow, Scotland, UK: ASRANet Ltd, 2008.

[39] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. 2nd ed. Springer Series in Statistics. New York: Springer, 2004. DOI: 10.1007/978-1-4757-4145-2.

[40] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[41] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97.

[42] M. K. Cowles and B. P. Carlin. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review". In: *Journal of the American Statistical Association* 91.434 (1996), pp. 883–904. DOI: 10.1080/01621459.1996.10476956.

[43] S. P. Brooks and G. O. Roberts. "Convergence assessment techniques for Markov chain Monte Carlo". In: *Statistics and Computing* 8.4 (1998), pp. 319–335. DOI: 10.1023/A:1008820505350.

[44] P. J. Diggle and R. J. Gratton. "Monte Carlo Methods of Inference for Implicit Statistical Models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2 (1984), pp. 193–227.

[45] P. D. O'Neill, D. J. Balding, N. G. Becker, M. Eerola, and D. Mollison. "Analyses of Infectious Disease Data from Household Outbreaks by Markov chain Monte Carlo Methods". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49.4 (2000), pp. 517–542. DOI: 10.1111/1467-9876.00210.

[46] M. A. Beaumont. "Estimation of Population Growth or Decline in Genetically Monitored Populations". In: *Genetics* 164.3 (2003), pp. 1139–1160.

[47] G. Bal, I. Langmore, and Y. Marzouk. "Bayesian Inverse Problems with Monte Carlo Forward Models". In: *Inverse Problems and Imaging* 7.1 (2013), pp. 81–105. DOI: 10.3934/ipi.2013.7.81.

[48] A. Korattikara, Y. Chen, and M. Welling. "Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget". In: *JMLR Workshop and Conference Proceedings: 31st International Conference on Machine Learning (ICML 2014)* 32.1 (2014), pp. 181–189.

[49] R. Bardenet, A. Doucet, and C. Holmes. "Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach". In: *JMLR Workshop and Conference Proceedings: 31st International Conference on Machine Learning (ICML 2014)* 32.1 (2014), pp. 405–413.

[50] J. M. McFarland, B. J. Bichon, and D. S. Riha. "A Probabilistic Treatment of Multiple Uncertainty Types: NASA UQ Challenge". In: *16th AIAA Non-Deterministic Approaches Conference (SciTech 2014)*. Reston, Virginia, USA: American Institute of Aeronautics and Astronautics (AIAA), 2014. DOI: 10.2514/6.2014-1500.

[51] C. Andrieu and G. O. Roberts. "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations". In: *The Annals of Statistics* 37.2 (2009), pp. 697–725. DOI: 10.1214/07-AOS574.

[52] J. C. Helton and W. L. Oberkampf. "Alternative representations of epistemic uncertainty". In: *Reliability Engineering & System Safety* 85.1–3 (2004), pp. 1–10. DOI: 10.1016/j.ress.2004.03.001.

[53] J. C. Helton and J. D. Johnson. "Quantification of margins and uncertainties: Alternative representations of epistemic uncertainty". In: *Reliability Engineering & System Safety* 96.9 (2011), pp. 1034–1052. DOI: 10.1016/j.ress.2011.02.013.

[54] M. A. Tanner and W. H. Wong. "The Calculation of Posterior Distributions by Data Augmentation". In: *Journal of the American Statistical Association* 82.398 (1987), pp. 528–540. DOI: 10.1080/01621459.1987.10478458.

[55] D. A. van Dyk and X.-L. Meng. "The Art of Data Augmentation". In: *Journal of Computational and Graphical Statistics* 10.1 (2001), pp. 1–50. DOI: 10.1198/10618600152418584.

[56] D. A. van Dyk. "Hierarchical Models, Data Augmentation, and Markov Chain Monte Carlo". In: *Statistical Challenges in Astronomy*. New York: Springer, 2003, pp. 41–55. DOI: 10.1007/0-387-21529-8_3.

[57] J. B. Nagel and B. Sudret. "A unified framework for multilevel uncertainty quantification in Bayesian inverse problems". In: *Probabilistic Engineering Mechanics* 43 (2016), pp. 68–84. DOI: 10.1016/j.probengmech.2015.09.007.

Chapter 7

Bayesian assessment of structural masonry

Original publication

J. B. Nagel, N. Mojsilovic, and B. Sudret. "Bayesian Assessment of the Compressive Strength of Structural Masonry". In: *12th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP12)*. Vancouver, Canada: University of British Columbia, 2015. DOI: 10.14288/1.0076072

Abstract

The application of hierarchical models for assessing the compressive strength of structural masonry is investigated. Based on current codified models the distribution of compressive strengths within an ensemble of masonry wall specimens is related to the statistical properties of the populations of brick units and mortar used. The parameters of this relation are calibrated with test data acquired at ETH Zürich. This approach allows for heterogeneous material modeling, consistent uncertainty management and optimal information processing. Costly compression tests of full-size masonry and inexpensive tests of brick and mortar samples are jointly utilized for learning about the masonry wall characteristics.

7.1 Introduction

Structural masonry is a composite material that consists of brick units and mortar. A simplified sketch is found in Fig. 7.1. The mechanical key characteristic of masonry is the compressive strength perpendicular to the bed joints. Estimating or predicting this material property are thus issues of central importance to assessing the reliability of masonry structures. These problems are therefore addressed in current standards [1, 2] and numerous enhancements [3–10].

The motivation of this research study is twofold. Firstly, we observe a systematic discrepancy between measured data and predictions of the masonry compressive strength according to [1]. This suggests a recalibration of the model code parameters. Secondly, it is noticed that current approaches either suffer from their semi-probabilistic character or their unsatisfactory treatment of the emerging uncertainties. Thus the goal of this paper is to develop a fully probabilistic extension of current codified models for assessing the compressive strength of unreinforced masonry. We will rely on hierarchical models [11, 12] and Bayesian networks [13, 14].

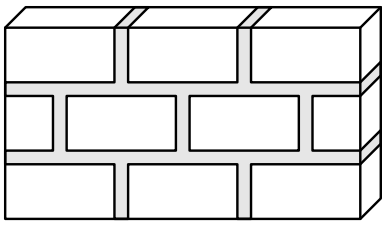


Figure 7.1: Structural masonry. The masonry wall is composed of brick units (white) that are bound by mortar (gray).

This approach will allow for heterogeneous modeling of structural masonry, quantification of various types of uncertainty and acquisition of information from diverse sources.

More specifically it is aimed at analyzing the compressive strength of structural masonry with system-level data, i.e. measurements that are taken from full-scale masonry specimens, component-level data, i.e. results from testing brick units and mortar samples individually, and prior or expert knowledge. Compression tests of masonry specimens are rather costly, whereas data associated to component-specific material characteristics are relatively inexpensive to acquire. Hierarchical models enable the joint processing of information from different levels of the overall system. This way the information is optimally utilized. Moreover a predictive relationship is established that connects the masonry compressive strength with the component-level compressive strengths.

The remainder of this document is organized as follows. Previous approaches of assessing the compressive strength of structural masonry will be reviewed in Section 7.2. Hierarchical models will be introduced in Section 7.3. In Section 7.4 the acquired data will be discussed and Section 7.5 will show the results of Bayesian updating. Lastly we will summarize and conclude in Section 7.6.

7.2 Current models

In [1] it is tried to relate the compressive strength of masonry to the resistances of its brick and mortar components. The relationship is realized as a power function

$$f_w = k' f_b^{\alpha'} f_m^{\beta'}. \tag{7.1}$$

On the one hand, the compressive strength of masonry is summarized by the characteristic value f_w, i.e. a 5 %-quantile. On the other hand, f_b denotes the normalized mean compressive strength of the units and f_m denotes the mean compressive strength of the mortar. Estimates of the constants (k', α', β') are given for different types of masonry. In [2] the empirical relation Eq. (7.1) is interpreted similarly. Here f_w, f_b and f_m represent the mean values of the corresponding distributions. Different prior estimates of the coefficients (k', α', β') are provided. The coefficients are often set so that they (approximately) satisfy $\alpha' + \beta' = 1$. This choice can be justified for reasons of the physical dimension in Eq. (7.1).

Ensuing from these semi-probabilistic models, a variety of extensions have been proposed in the literature. There are probabilistic reinterpretations of Eq. (7.1) based on lognormal distributions [5, 7, 9, 10]. In other studies the model uncertainty of Eq. (7.1) is quantified [3, 4]. A conjugate Bayesian updating approach based on Gaussian distributions is presented in [6]. Another idea is to establish a connection between the compressive strengths of masonry and its components via artificial neural networks [8].

These previous approaches suffer from the fact that they either do not clearly distinguish between epistemic and aleatory shares of uncertainty or they neglect material heterogeneity. Fitting the parameters of a probabilistic extension of Eq. (7.1) is a problem that has hardly been satisfactorily solved as yet.

7.3 Hierarchical models

In the following hierarchical Bayesian modeling is introduced as a tool for distinguishing and handling uncertainty in codified models of the form Eq. (7.1). The aim of this section is to establish a Bayesian model and updating strategy for the following experimental situation. The compressive strength is measured for a number of clay block masonry specimens. Specimens can be grouped according to the ensembles of brick units and mortar that were used for their construction. Here ensembles of clay bricks are characterized by the same ingredients used and the same manufacturing procedure. Similarly in every ensemble of mortar samples identical constituents were used for mixing. In this modeling approach material heterogeneity is accounted for by distinguishing between brick and mortar samples used in constructing the masonry wall systems. The final goal is the assessment and prediction of the compressive capacity of structural masonry by utilizing system- and component level information.

7.3.1 Aleatory model

Within an ensemble of masonry wall specimens, the compressive strength of the masonry wall is represented as a random variable

$$F_w = k F_b^{\alpha} F_m^{\beta}. \tag{7.2}$$

This is a probabilistic extension of the codified model in Eq. (7.1). We remark that the coefficients (k, α, β) of the relation Eq. (7.2) are not immediately identified with the ones of Eq. (7.1).

The compressive strengths of the bricks and the mortar are modeled as lognormal random variables $F_b \sim \mathcal{LN}(f_b|\mu_b, \sigma_b^2)$ and $F_m \sim \mathcal{LN}(f_m|\mu_m, \sigma_m^2)$. Their distributions are determined by hyperparameters $\boldsymbol{\theta}_b = (\mu_b, \sigma_b)$ and $\boldsymbol{\theta}_m = (\mu_m, \sigma_m)$ that are the mean and standard deviation of $\log(F_b)$ and $\log(F_m)$, respectively. Consequently the masonry wall compressive strength in Eq. (7.2) is a random variable that follows a lognormal distribution

$$F_w \sim \mathcal{LN}\left(f_w|\mu_w, \sigma_w^2\right), \tag{7.3a}$$

$$\text{with } \mu_w = \alpha\mu_b + \beta\mu_m + \log k, \tag{7.3b}$$

$$\text{and } \sigma_w^2 = \alpha^2\sigma_b^2 + \beta^2\sigma_m^2. \tag{7.3c}$$

The distribution Eq. (7.3a) represents the variability, i.e. the frequency distribution, of the masonry compressive strengths within the population of specimens. It is parametrized by hyperparameters $\boldsymbol{\theta}_w = (\mu_w, \sigma_w)$ that are determined by the statistical properties of component populations due to Eqs. (7.3b) and (7.3c).

The mean value and the variance of the distribution $\mathcal{LN}\left(f_w|\mu_w, \sigma_w^2\right)$ in Eq. (7.3) are simply given as $\mathbb{E}[F_w] = \exp(\mu_w + \sigma_w^2/2)$ and $\mathrm{Var}[F_w] = (\exp(\sigma_w^2) - 1)\exp(2\mu_w + \sigma_w^2)$, respectively. The 5 %-quantile of $\mathcal{LN}\left(f_w|\mu_w, \sigma_w^2\right)$, e.g. for comparison with Eq. (7.1), follows as $Q_{w,5\,\%} = \exp(\mu_w - 1.645\sigma_w)$.

7.3.2 Epistemic model

If the coefficients (k, α, β) of Eqs. (7.2) and (7.3) are not perfectly known, one can represent their epistemic uncertainty as prior random variables $(K, A, B) \sim \pi(k, \alpha, \beta)$. In the following we will confine the analysis to the case $\beta = 1 - \alpha$. We consider mutually independent prior random variables

$$K \sim \pi(k), \quad A \sim \pi(\alpha). \tag{7.4}$$

Their joint prior uncertainty $\pi(k, \alpha) = \pi(k)\,\pi(\alpha)$ can be reduced by Bayesian data analysis of experimental measurements. In the following two different updating approaches are outlined for experimental situations where the assumption of known hyperparameters, i.e. the distributional parameters of the ensembles of masonry wall components, is either justified or rather unfounded.

7.3.3 Known hyperparameters

Let us consider experiments of the following type. In each batch of experiments $i = 1, \ldots, n$ the masonry compressive strength $f_{w,ij}$ is measured for a number of different specimens $j = 1, \ldots, J_i$ from an ensemble. We use $\langle f_{w,ij}\rangle = (f_{w,11}, \ldots, f_{w,nJ_n})$ to denote the set of these measurements. The hyperparameters $\boldsymbol{\theta}_{b,i}$ and $\boldsymbol{\theta}_{m,i}$ are measured for the bricks and the mortar used in experiment i, too. This can be accomplished by a statistical analysis of data $\langle f_{b,ik}\rangle = (f_{b,11}, \ldots, f_{b,nK_n})$ and $\langle f_{m,il}\rangle = (f_{m,11}, \ldots, f_{m,nK_n})$ with $k = 1, \ldots, K_i$ and $l = 1, \ldots, L_i$. These data must be numerous and they must be observed for the ensembles of brick units and mortar used. The Bayesian multilevel model for this scenario can be written as

$$(F_{w,ij}|k, \alpha) \sim \pi(f_{w,ij}|k, \alpha), \tag{7.5a}$$

$$(K, A) \sim \pi(k)\,\pi(\alpha). \tag{7.5b}$$

Here the conditional distributions Eq. (7.5a) are given by Eq. (7.3), where batch-specific knowns $\boldsymbol{\theta}_{b,i}$ and $\boldsymbol{\theta}_{b,i}$ are plugged in. The epistemic prior uncertainty of the coefficients (k, α) is encoded in Eq. (7.5b). As long as not indicated otherwise, all random variables in Eq. (7.5) are assumed to be (conditionally) independent. A directed acyclic graph (DAG) as in Fig. 7.2 serves as an intuitive visualization of the model Eq. (7.5).

As usual, Bayesian updating is accomplished by conditioning the prior distribution $\pi(k, \alpha) = \pi(k)\,\pi(\alpha)$ on the acquired data $\langle f_{w,ij}\rangle$. One obtains

$$\pi(k, \alpha|\langle f_{w,ij}\rangle) \propto \pi(k)\,\pi(\alpha) \prod_{i=1}^{n}\prod_{j=1}^{J_i} \pi(f_{w,ij}|k, \alpha). \tag{7.6}$$

Note that Eq. (7.6) is based on exact values the hyperparameters $\boldsymbol{\theta}_{b,i}$ and $\boldsymbol{\theta}_{m,i}$ for every batch i.

7.3.4 Unknown hyperparameters

The requirement of known hyperparameters $\boldsymbol{\theta}_{b,i}$ and $\boldsymbol{\theta}_{m,i}$ restricts the applicability model Eq. (7.5) to situations that are rarely met in practice. Therefore we consider the situation when only prior knowledge $\pi(\boldsymbol{\theta}_{b,i}, \boldsymbol{\theta}_{m,i}) = \pi(\boldsymbol{\theta}_{b,i})\,\pi(\boldsymbol{\theta}_{m,i})$ about the hyperparameters is available. Additionally in each batch of experiments

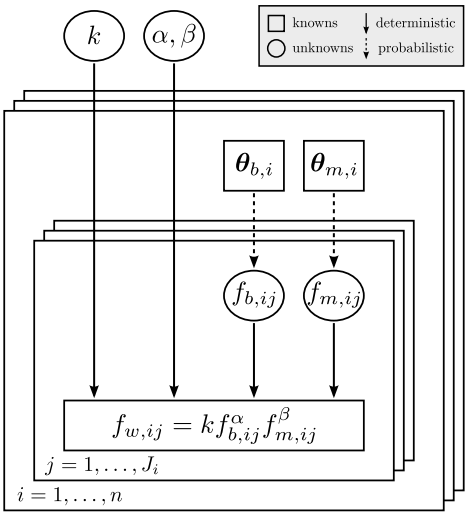


Figure 7.2: Known hyperparameters. Nodes symbolize known (□) or unknown (○) quantities. Arrows represent deterministic (⟶) or probabilistic (⇢) relations.

i a variable number of measurements $f_{b,ik}$ and $f_{m,il}$ for $k = 1, \ldots, K_i$ and $l = 1, \ldots, L_i$ are taken of the brick unit and the mortar compressive strength, respectively. The corresponding hierarchical Bayesian model reads

$$\begin{aligned} (F_{w,ij} | k, \alpha, \boldsymbol{\theta}_{b,i}, \boldsymbol{\theta}_{m,i}) &\sim \pi(f_{w,ij} | k, \alpha, \boldsymbol{\theta}_{b,i}, \boldsymbol{\theta}_{m,i}), \\ (F_{b,ik} | \boldsymbol{\theta}_{b,i}) &\sim \pi(f_{b,ik} | \boldsymbol{\theta}_{b,i}), \\ (F_{m,il} | \boldsymbol{\theta}_{m,i}) &\sim \pi(f_{m,il} | \boldsymbol{\theta}_{m,i}), \end{aligned} \tag{7.7a}$$

$$\begin{aligned} (\boldsymbol{\Theta}_{b,i}, \boldsymbol{\Theta}_{m,i}) &\sim \pi(\boldsymbol{\theta}_{b,i})\, \pi(\boldsymbol{\theta}_{m,i}), \\ (K, A) &\sim \pi(k)\, \pi(\alpha). \end{aligned} \tag{7.7b}$$

While Eq. (7.7a) summarizes the aleatory uncertainties, Eq. (7.7b) contains the epistemic uncertainties. The model Eq. (7.7) is visualized as the DAG in Fig. 7.3. We remark that the observations $\langle f_{b,ik} \rangle$ and $\langle f_{m,il} \rangle$ inform about the statistical properties $\boldsymbol{\theta}_{b,i}$ and $\boldsymbol{\theta}_{m,i}$ of the component ensembles. This way they give information about the unobservable properties of the brick and mortar samples used for constructing the masonry wall i.

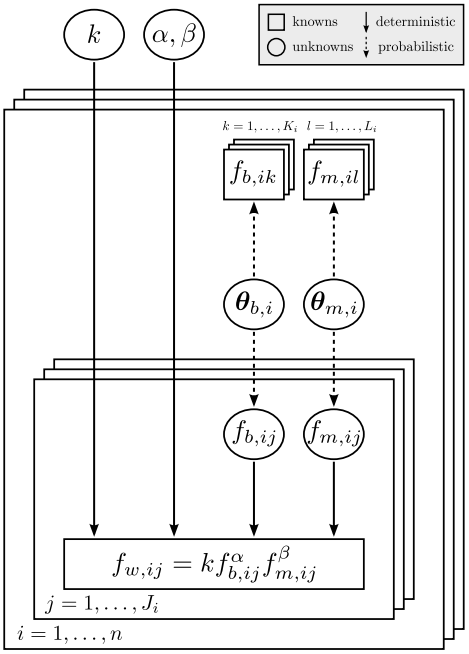


Figure 7.3: Unknown hyperparameters. The batch-specific hyperparameters $\boldsymbol{\theta}_{b,i}$ and $\boldsymbol{\theta}_{m,i}$ are unknown. They can be inferred from the data $\langle f_{b,ik} \rangle$ and $\langle f_{m,il} \rangle$.

Bayesian analysis proceeds by updating the joint prior $\pi(k, \alpha, \langle \boldsymbol{\theta}_{b,i} \rangle, \langle \boldsymbol{\theta}_{m,i} \rangle) = \pi(k)\,\pi(\alpha)\,\prod_{i=1}^{n} \pi(\boldsymbol{\theta}_{b,i})\,\pi(\boldsymbol{\theta}_{m,i})$. Conditioned on the data $(\langle f_{w,ij} \rangle, \langle f_{b,ik} \rangle, \langle f_{m,il} \rangle)$ one obtains for the joint posterior

$$\begin{aligned} \pi(k, \alpha, \langle \boldsymbol{\theta}_{b,i} \rangle, \langle \boldsymbol{\theta}_{m,i} \rangle | \langle f_{w,ij} \rangle, \langle f_{b,ik} \rangle, \langle f_{m,il} \rangle) \propto \pi(k)\,\pi(\alpha) \prod_{i=1}^{n} \pi(\boldsymbol{\theta}_{b,i})\,\pi(\boldsymbol{\theta}_{m,i}) \\ \cdot \prod_{j=1}^{J_i} \pi(f_{w,ij} | k, \alpha, \boldsymbol{\theta}_{b,i}, \boldsymbol{\theta}_{m,i}) \prod_{k=1}^{K_i} \pi(f_{b,ik} | \boldsymbol{\theta}_{b,i}) \prod_{l=1}^{L_i} \pi(f_{m,il} | \boldsymbol{\theta}_{m,i}). \end{aligned} \tag{7.8}$$

Notice that the posterior Eq. (7.8) gathers information from both system- and component-level data.

7.4 Experimental data

In the years 2009-2012 and 2014 the compressive strength of clay block masonry was measured for a variable number of specimens in a series of compression tests. In addition, the compressive strengths of bricks and mortar were recorded for realizations from the same ensembles that were later used for the construction of the masonry wall. The tests were performed at the laboratories of the Department of Civil, Environmental and Geomatic Engineering of ETH Zürich. Two photographs that were made during the tests are shown in Fig. 7.4. In Table 7.1 the experimental data are summarized. Five batches of experiments were performed in total. At the system- and the component level the available data is generally scarce. Especially in the years 2011 and 2012 the number of component-level tests was very limited. Moreover, in the years 2009 and 2010 brick units from the same ensemble were used.

(a) Before failure.

(b) After failure.

Figure 7.4: Compression test. A specimen is shown before its failure in (a) and thereafter in (b).

We observe that the empirical relation Eq. (7.1) generally overpredicts the masonry wall compressive strength. In Fig. 7.5 the actually acquired data for $i = 1$, i.e. for the year 2009, is shown together with the correspondingly predicted characteristic value. The values $k' = 0.45$, $\alpha' = 0.7$, $\beta' = 0.3$ provided in [1] were used. Brick unit data $\langle f_{b,ik} \rangle$ have been normalized according to their geometry. Moreover a lognormal distribution of the form Eq. (7.2) is shown, where $\alpha = \alpha'$ and $\beta = \beta'$ have been identified with the corresponding coefficients from [1]. The remaining coefficient $k = k' \cdot \exp(1.645\sqrt{\alpha_i^2 \sigma_{b,i}^2 + \beta_i^2 \sigma_{m,i}^2} + \alpha_i[\sigma_{b,i}^2/2] + \beta_i[\sigma_{m,i}^2/2])$ has been set in order that the 5 %-quantile equals Eq. (7.1). Note that the abovementioned identification/transformation of the coefficients establishes another way of extending Eq. (7.1) and comparing it to Eq. (7.2). In this paper we do not pursue this approach, though.

Of course, the unexpected code/measurement discrepancy raises important questions. Anticipating our results it is said that we will not be able to satisfactorily explain this discrepancy. Instead we will calibrate the coefficients k and α in a way that leads to better predictions. Those predictions are valid for the testing

Table 7.1: Experimental data. Data are shown for tests of clay block masonry that were performed in the years 2009-2012 and in 2014. Blocks from the same ensemble were used in 2009 and 2010. Thus the corresponding rows show duplicate data entries.

	2009	Batch 1												
f_w	[MPa]	9.41	5.53	7.98	8.86	6.67	7.17	7.92	-	-	-	-	-	-
f_b	[MPa]	33.58	34.55	37.1	39.21	39.63	36.1	35.46	37.61	35.6	36.26	35.2	32.7	36.6
f_m	[MPa]	15.8	16.1	14.4	14.8	16.1	15.4	13.9	14.6	14.6	14.4	16.8	16.1	-
	2010	Batch 2												
f_w	[MPa]	6.67	6.12	5.91	8.3	6.44	5.32	6.7	-	-	-	-	-	-
f_b	[MPa]	33.58	34.55	37.1	39.21	39.63	36.1	35.46	37.61	35.6	36.26	35.2	32.7	36.6
f_m	[MPa]	12.5	12.94	12.43	13.33	12	12.32	-	-	-	-	-	-	-
	2011	Batch 3												
f_w	[MPa]	4.32	3.71	6.06	4.95	4.29	2.8	6.28	4.22	5.23	-	-	-	-
f_b	[MPa]	23.8	26.8	25.7	-	-	-	-	-	-	-	-	-	-
f_m	[MPa]	14.9	14.7	14.9	15.4	14.7	14.6	-	-	-	-	-	-	-
	2012	Batch 4												
f_w	[MPa]	8	7.87	8.1	7.53	8.14	6.99	7.82	9.13	5.87	7.71	-	-	-
f_b	[MPa]	37	39.9	38	-	-	-	-	-	-	-	-	-	-
f_m	[MPa]	26.9	28.1	17	16.2	18.7	21.1	-	-	-	-	-	-	-
	2014	Batch 5												
f_w	[MPa]	6.53	7.01	6.12	5.94	7.14	5.69	5.82	6.34	5.96	-	-	-	-
f_b	[MPa]	28.15	27.74	28.05	27.20	26.25	23.15	26.69	28.04	27.69	26.73	-	-	-
f_m	[MPa]	11.73	12.19	12.13	10.49	10.34	10.44	-	-	-	-	-	-	-

machine and the materials used in our laboratory. Using the predictions outside their scope of applicability is questionable and should only be done with utmost caution.

7.5 Bayesian analysis

The Bayesian framework discussed in Section 7.3 is now applied to analyze the experimental data that was presented in Section 7.4. More specifically we use the first two batches of experiments that were conducted in 2009 and 2010 to calibrate the unknown coefficients of the model Eq. (7.5). For those batches the amount of component-level data is deemed sufficient to fit the hyperparameters and to treat them as knowns subsequently. Moreover the first four batches will be analyzed with the model Eq. (7.7) that allows to treat the hyperparameters as unknowns. Especially in the years 2011 and 2012 the small amount of component-level data does not allow to proceed in another way. The fifth batch of experiments from 2014 will be used as an independent test set.

Since the coefficients in Eqs. (7.2) and (7.3) cannot be identified with those of Eq. (7.1), it is not possible to elicit informative priors about the former by exploiting expert knowledge or code information about the latter. Hence uninformative priors are used. Specifically we assign uniform prior distributions $\pi(k) = \mathcal{U}(0,1)$ and $\pi(\alpha) = \mathcal{U}(0.5,1)$. Due to $\beta = 1-\alpha$ the latter assignment enforces $\alpha \geq \beta$. This reflects the intuition that, regarding the masonry compressive strength, the brick units are more influential than the mortar. For the Bayesian model in Eq. (7.7) priors $\pi(\boldsymbol{\theta}_{b,i}) = \pi(\mu_{b,i})\,\pi(\sigma_{b,i})$ and $\pi(\boldsymbol{\theta}_{m,i}) = \pi(\mu_{b,i})\,\pi(\sigma_{b,i})$ have to be elicited for the unknown hyperparameters. We use independent uniform hyperprior distributions with reasonable bounds for the means and standard deviations.

The posteriors Eqs. (7.6) and (7.8) can be sampled by means of Markov chain Monte Carlo (MCMC) techniques [15]. In Figs. 7.6 and 7.7 the resulting posterior marginals of k and α are depicted. It can be seen that $\pi(k,\alpha|\langle f_{w,ij}\rangle, \langle f_{b,ik}\rangle, \langle f_{m,il}\rangle)$ contains a higher degree of posterior uncertainty than $\pi(k,\alpha|\langle f_{w,ij}\rangle)$. Since more data has entered the former posterior, at first sight this seems to be surprising. This fact can be attributed to the differences of the models Eqs. (7.5) and (7.7) in treating the hyperparameters and their uncertainties, though.

Specifically the modes $\hat{k} = 0.21$ and $\hat{\alpha} = 1$ are found for the posterior $\pi(k,\alpha|\langle f_{w,ij}\rangle)$ that represents the situation that hyperparameters are assumed to be known. The posterior $\pi(k,\alpha|\langle f_{w,ij}\rangle, \langle f_{b,ik}\rangle, \langle f_{m,il}\rangle)$, for the

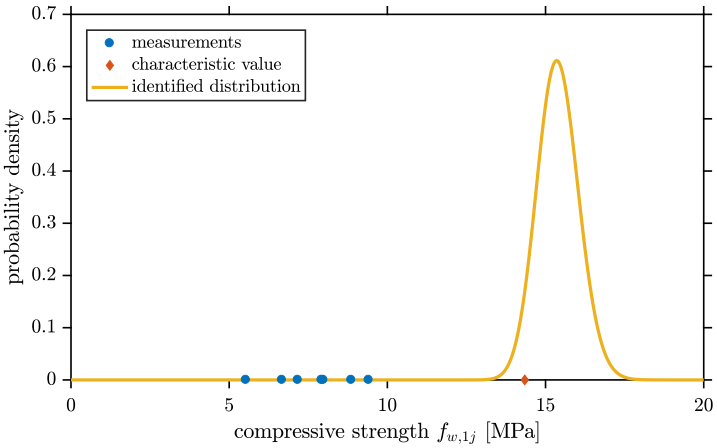


Figure 7.5: Data and predictions for 2009. The data, its expected 5 %-quantile and a corresponding lognormal distribution are shown. The data are overpredicted.

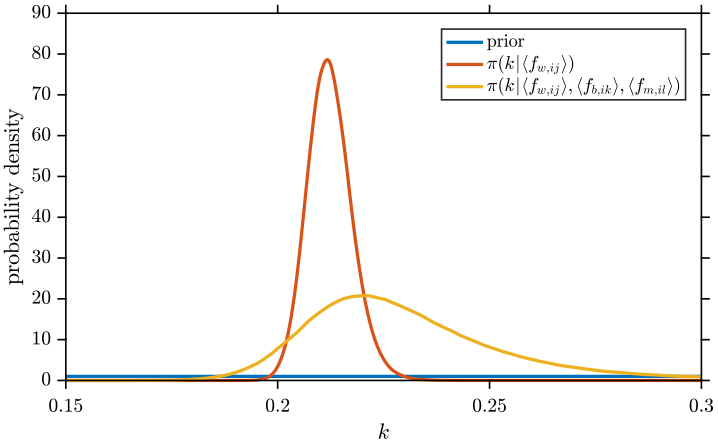


Figure 7.6: Posterior of k. The posteriors $\pi(k|\langle f_{w,ij}\rangle)$ and $\pi(k|\langle f_{w,ij}\rangle, \langle f_{b,ik}\rangle, \langle f_{m,il}\rangle)$ are shown. It can be seen that the latter is broader than the former.

scenario that hyperparameters are treated as unknowns, features the modes $\hat{k} = 0.22$ and $\hat{\alpha} = 1$.

The fact that the posterior of α in Fig. 7.7 peaks at the upper bound of its prior is somewhat surprising. As a consequence of $\hat{\beta} = 1 - \hat{\alpha} = 0$, the influence of mortar occurs to be negligible. Moreover, such a behavior may indicate that the inverse problem is improperly solved, e.g. the true parameter value was accidentally excluded a priori. It was therefore tried to relax the assumption $\beta = 1 - \alpha$ by permitting arbitrary values $\alpha > 0$ and $\beta > 0$. To that end independent priors $\pi(\alpha)$ and $\pi(\beta)$ were assigned. We had to conclude that the limited amount of available data is not sufficiently informative in order to calibrate this extended model.

Plugging the point estimates $\hat{k}$ and $\hat{\alpha}$ in Eq. (7.3) establishes a predictive relation of the frequency distribution of structural masonry. For that purpose one has to specify the values or estimates of the hyperparameters $\boldsymbol{\theta}_b$ and $\boldsymbol{\theta}_m$ for the ensembles of bricks and mortar used in the construction of the masonry wall. The predicted distributions, that are obtained this way for the actually analyzed batches of experiments, describe the masonry wall resistances adequately well. Since the estimations of the coefficients were informed by the very same data, this does not seem to be very surprising. Yet this signifies that the representation Eq. (7.3) is adjustable enough to match the data. In turn this may indicate that Eq. (7.3) is indeed a suitable representation of the masonry wall compressive strength.

When applied to the fifth batch of experiments the procedure described above can serve as a validation test, i.e. the data collected in 2014 are used as an independent test set. In Fig. 7.8 the measured masonry wall compressive strengths are shown together with the their predicted distribution. The plot is supplemented with the corresponding 5 %-quantile. Here the point estimates $\hat{k} = 0.22$ and $\hat{\alpha} = 1$ that were obtained by analyzing the previous four batches are used on one side. On the other side component-level data $\langle f_{b,5j}\rangle$ and $\langle f_{m,5j}\rangle$ for the fifth batch are used to estimate $\boldsymbol{\theta}_{b,5}$ and $\boldsymbol{\theta}_{m,5}$. The predictive distribution captures the data fairly well. Obviously it is of higher quality than the poor code-forecast shown in Fig. 7.5.

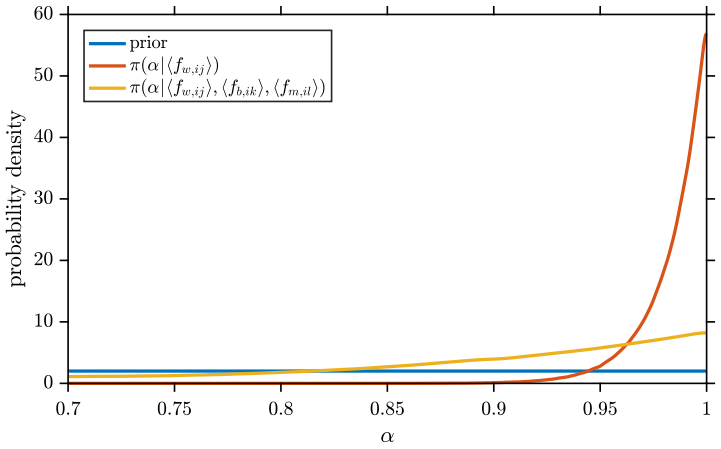


Figure 7.7: Posterior of α. Both the posterior marginals $\pi(\alpha|\langle f_{w,ij}\rangle)$ and $\pi(\alpha|\langle f_{w,ij}\rangle, \langle f_{b,ik}\rangle, \langle f_{m,il}\rangle)$ peak at their upper boundary.

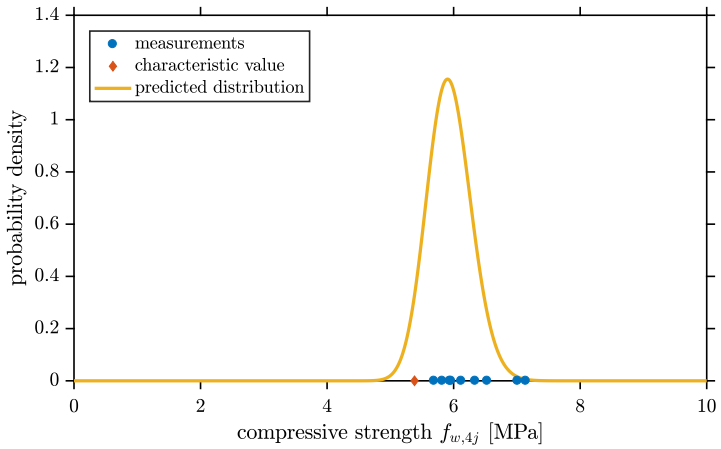


Figure 7.8: Data and predictions for 2014. The gathered data, the predicted distribution and its 5 %-quantile are shown. Predictions conform to data tolerably well.

7.6 Summary and conclusion

It was demonstrated how hierarchical Bayesian models can serve the purpose of assessing the compressive strength of structural masonry. This establishes a fully probabilistic alternative to the existing semi-probabilistic approaches. The hierarchical framework offers versatile and powerful tools of uncertainty quantification and information aggregation at multiple system levels. Different types of uncertainty, i.e. ignorance and variability, are thoroughly managed, while heterogeneous types of information, e.g. data and expert knowledge, are consistently utilized. This way the analysis of the masonry wall resistance can be based on large-scale compression tests as well as on inexpensive tests of brick unit and mortar samples.

Our hope is that this possibility will encourage experimenters in entirely publishing their collected data. In fact it seems to be commonplace to quote statistical data summaries only, e.g. sample means or characteristic values. The proposed methodology, however, allows to process the acquired data as a whole.

A number of questions have arisen. It is queried if Eq. (7.3) is an adequate representation of the distribution of masonry compressive strength in terms of distributional parameters of the components. With regard to the complexity of structural masonry, its failure modes and their dependency on the quality of workmanship, the relations Eqs. (7.1) and (7.2) are oversimplifying. They were inspired by the structure of current models but lack a solid physical foundation. For future studies this motivates the introduction of model uncertainty in addition to the emerging parameters uncertainties. Beyond that future work will also involve the construction and objective selection of better system-level models of aleatory variability. A more fundamental question concerns the general suitability of empirical relations for any probabilistic extension whatsoever. Another raised issue relates to the observed mismatch between measurements and code-predictions. We were not able to explain this discrepancy.

References

[1] *Eurocode 6: Design of masonry structures. Part 1-1: General rules for reinforced and unreinforced masonry structures*. Brussels, Belgium: European Committee for Standardization (CEN), 2005.

[2] *Probabilistic Model Code*. Zürich, Switzerland: Joint Committee on Structural Safety (JCSS), 2001.

[3] C. Dymiotis and B. M. Gutlederer. "Allowing for uncertainties in the modelling of masonry compressive strength". In: *Construction and Building Materials* 16.8 (2002), pp. 443–452. DOI: 10.1016/S0950-0618(02)00108-3.

[4] S. Glowienka and C.-A. Graubner. "Probabilistic Modelling of the Load Carrying Capacity of Modern Masonry". In: *4th International Probabilistic Symposium*. Berlin, Germany, October 2006.

[5] L. Schueremans and D. Van Gemert. "Probability density functions for masonry material parameters – a way to go ?" In: *5th International Conference on Structural Analysis of Historical Constructions: Possibilities of Numerical and Experimental Techniques*. New Delhi, India, November 2006.

[6] N. Mojsilovic and M. H. Faber. "Probabilistic assessment of masonry compressive strength". In: *10th International Conference on Structural Safety and Reliability (ICOSSAR2009)*. Osaka, Japan, September 2009.

[7] M. Sýkora and M. Holický. "Probabilistic Model for Masonry Strength of Existing Structures". In: *Engineering Mechanics* 17.1 (2010), pp. 61–70.

[8] J. Garzón-Roca, C. Obrer Marco, and J. M. Adam. "Compressive strength of masonry made of clay bricks and cement mortar: Estimation based on Neural Networks and Fuzzy Logic". In: *Engineering Structures* 48 (2013), pp. 21–27. DOI: 10.1016/j.engstruct.2012.09.029.

[9] M. Sykora and M. Holicky. "Evaluation of Compressive Strength of Historic Masonry Using Measurements". In: *Advanced Materials Research* 923 (2014), pp. 213–216. DOI: 10.4028/www.scientific.net/AMR.923.213.

[10] M. Sykora, T. Cejka, M. Holicky, and J. Witzany. "Probabilistic model for compressive strength of historic masonry". In: *European Safety and Reliability Conference (ESREL 2013)*. Ed. by R. D. J. M. Steenbergen, P. H. A. J. M. van Gelder, S. Miraglia, and A. C. W. M. Ton Vrouwenvelder. Leiden, Netherlands: CRC Press/Balkema, 2014. Chap. 315, pp. 2645–2652. DOI: 10.1201/b15938-400.

[11] J. B. Nagel and B. Sudret. "Bayesian Multilevel Model Calibration for Inverse Problems Under Uncertainty with Perfect Data". In: *Journal of Aerospace Information Systems* 12.1 (2015), pp. 97–113. DOI: 10.2514/1.I010264.

[12] J. B. Nagel and B. Sudret. "A unified framework for multilevel uncertainty quantification in Bayesian inverse problems". In: *Probabilistic Engineering Mechanics* 43 (2016), pp. 68–84. DOI: 10.1016/j.probengmech.2015.09.007.

[13] S. Sankararaman, K. McLemore, and S. Mahadevan. "Bayesian Methods for Uncertainty Quantification in Multi-level Systems". In: *Topics in Model Validation and Uncertainty Quantification, Volume 4*. Ed. by T. Simmermacher, S. Cogan, L. G. Horta, and R. Barthorpe. Conference Proceedings of the Society for Experimental Mechanics Series. New York: Springer, 2012, pp. 67–74. DOI: 10.1007/978-1-4614-2431-4_7.

[14] A. Urbina, S. Mahadevan, and T. L. Paez. "A Bayes Network Approach to Uncertainty Quantification in Hierarchically Developed Computational Models". In: *International Journal for Uncertainty Quantification* 2.2 (2012), pp. 173–193. DOI: 10.1615/Int.J.UncertaintyQuantification.v2.i2.70.

[15] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds. *Handbook of Markov Chain Monte Carlo*. Handbooks of Modern Statistical Methods. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2011. DOI: 10.1201/b10905.

Chapter 8

Spectral likelihood expansions for Bayesian inference

Original publication

J. B. Nagel and B. Sudret. "Spectral likelihood expansions for Bayesian inference". In: *Journal of Computational Physics* 309 (2016), pp. 267–294. DOI: 10.1016/j.jcp.2015.12.047

Abstract

A spectral approach to Bayesian inference is presented. It pursues the emulation of the posterior probability density. The starting point is a series expansion of the likelihood function in terms of orthogonal polynomials. From this spectral likelihood expansion all statistical quantities of interest can be calculated semi-analytically. The posterior is formally represented as the product of a reference density and a linear combination of polynomial basis functions. Both the model evidence and the posterior moments are related to the expansion coefficients. This formulation avoids Markov chain Monte Carlo simulation and allows one to make use of linear least squares instead. The pros and cons of spectral Bayesian inference are discussed and demonstrated on the basis of simple applications from classical statistics and inverse modeling.

8.1 Introduction

In view of inverse modeling [1, 2] and uncertainty quantification [3, 4], Bayesian inference establishes a convenient framework for the data analysis of engineering systems [5–7]. It adopts probability theory in order to represent, propagate and update epistemic parameter uncertainties. The prior distribution captures the uncertainty of unknown model parameters before the data are analyzed. A posterior is then constructed as an updated distribution that captures the remaining uncertainty after the data have been processed. The computation of this posterior is the primary task in Bayesian inference.

Some simplified statistical models admit closed-form expressions of the posterior density. Beyond these so-called conjugate cases, computational approaches either aim at evaluating expectation values under the posterior or drawing samples from it [8]. This is usually accomplished with stochastic methods such as Markov chain Monte Carlo [9, 10]. Nowadays this class of techniques constitutes the mainstay of Bayesian computations. The posterior is explored by realizing an appropriate Markov chain over the prior support that exhibits the posterior as its long-run distribution. In turn, the obtained sample is used to empirically approximate the statistical quantities of interest. These include the characteristics of the posterior and the predictive distribution. This way of proceeding suffers from some inherent deficiencies. The presence of sample autocorrelation and the absence of a convergence criterion cause severe practical problems. Moreover, Markov chain Monte Carlo typically requires a large number of serial forward model runs. Since in engineering applications even a single model run can be computationally taxing, this may be prohibitive. In the recent past, numerous enhancements have been proposed in order to accelerate Markov chain Monte Carlo for Bayesian inverse problems. This includes the implementation of more efficient sampling algorithms, e.g. transitional Markov chain Monte Carlo [11, 12] or Hamiltonian Monte Carlo [13–15], and the substitution of the forward model with an inexpensive metamodel, e.g. based on Gaussian process models [16, 17] or polynomial chaos expansions [18–20]. Although these approaches promise significant speedups, they still inherit all principle shortcomings of sample-based posterior representations.

Unfortunately there are only few fundamental alternatives to stochastic sampling. Variational Bayesian inference establishes such an alternative where the posterior is sought through deterministic optimization [21–23]. In particular, a member from a simple parametric family of probability densities is selected such that some distance to the posterior is minimized. In this regard, the Kullback-Leibler divergence is often chosen as a relative measure of the dissimilarity of two probability densities. The procedure commonly rests upon some simplifying independence assumptions. Variational methods are regarded as less computing intensive than Markov chain Monte Carlo, yet they are only approximate. They are prominently used in machine learning and computer science [24, 25], and since recently such methods are applied to inverse problems [26, 27], too. A particularly interesting implementation of variational Bayesian inference has been proposed in [28]. The posterior is parametrized as a transformation of the prior density and can be computed based on the corresponding back-transformation. More specifically, a random variable transformation is sought in polynomial form such that the Kullback-Leibler divergence between the prior density and the back-transformed posterior density is minimized. This formulation is supported by arguments from optimal transport theory which also allows for a practical regularization of the problem. Finally, samples from the posterior distribution are obtained by independently sampling from the prior and applying the polynomial map. Another approach to certain Bayesian inverse problems has been recently devised in [29, 30]. Based on monomial Taylor expansions of the forward model and of the posterior density, the computation of expectation values under the posterior is tackled by sparse numerical quadrature.

In this paper we propose a novel approach to surrogate the posterior probability density in itself. The main idea is to decompose the likelihood function into a series of polynomials that are orthogonal with respect to the prior density. It is shown that all statistical quantities of interest can then be easily extracted from this spectral likelihood expansion. Emulators of the joint posterior density and its marginals are derived as the product of the prior, that functions as the reference density, and a linear combination of polynomials, that acts as an adjustment. In doing so, the model evidence simply emerges as the coefficient of the constant term in the expansion. Moreover, closed-form expressions for the first posterior moments in terms of the low-order expansion coefficients are given. The propagation of the posterior uncertainty through physical models can be easily accomplished based on a further postprocessing of the expansion coefficients. In this sense, spectral Bayesian inference is semi-analytic. While the corrections required for an expansion of the posterior with respect to the prior as the reference density may be large, they can be small for an expansion around a properly chosen auxiliary density. A change of the reference density is therefore suggested in order to increase the efficiency of computing a posterior surrogate. The devised formulation entirely avoids Markov chain Monte Carlo. Instead it draws on the machinery of spectral methods [31–33] and approximation theory [34–36]. It is proposed to compute the expansion coefficients via linear least squares [37, 38]. This allows one to make use of the wealth of statistical learning methods [39, 40] that are designed for this type of problems. The approach features a natural convergence criterion and it is amenable to parallel computing.

The scope of applicability of the proposed approach covers problems from Bayesian inference for which the likelihood function can be evaluated and for which polynomials can be constructed that are orthogonal with respect to the prior, possibly after a carefully chosen variable transformation. This excludes statistical models that involve intractable likelihoods [41, 42], i.e. the likelihood cannot be exactly evaluated. It also excludes improper prior distributions [43, 44], i.e. the prior does not integrate to one or any finite value, and models with pathologic priors such as the Cauchy distribution for which the moments are not defined [45, 46]. Many hierarchical Bayesian models [47, 48] are not covered by the devised problem formulation. They are either based on conditional priors, which does not allow for orthogonal polynomials, or on integrated likelihoods, which can only be evaluated subject to noise.

Spectral likelihood expansions complement the existing array of Bayesian methods with a way of surrogating the posterior density directly. They have the potential to remedy at least some of the shortcomings of Markov chain Monte Carlo. Yet, their practical implementation poses challenges. Hence, the goal of this paper is to discuss and investigate the possibilities and limitations of the approach. The method of spectral likelihood expansions is therefore applied to well-known calibration problems from classical statistics and inverse heat conduction. We restrict the analysis to low-dimensional problems. The final results are compared with corresponding results from Markov chain Monte Carlo simulations.

The manuscript is structured as follows. The principles of Bayesian inference are summarized in Section 8.2. Surrogate forward modeling with polynomial chaos expansions is reviewed in Section 8.3. After that, spectral likelihood expansions are introduced as an alternative approach to Bayesian inference in Section 8.4. Two well-known Gaussian fitting examples and the identification of thermal properties of a composite material serve as numerical demonstrations in Section 8.5. Finally, it is summarized and concluded in Section 8.6.

8.2 Bayesian inference

Let $\boldsymbol{x} = (x_1, \ldots, x_M)^\top \in \mathcal{D}_{\boldsymbol{x}}$ with $\mathcal{D}_{\boldsymbol{x}} = \mathcal{D}_{x_1} \times \ldots \times \mathcal{D}_{x_M} \subset \mathbb{R}^M$ be a vector of unknown parameters. The goal of statistical inference is to deduce these unknowns from the observed data $\boldsymbol{y} = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$. In Bayesian statistics one adapts probabilistic models for representing uncertainty. Hence, let $(\Omega, \mathcal{F}, \mathcal{P})$ be a suitable probability space with a sample space Ω, a σ-field $\mathcal{F}$ and a probability measure $\mathcal{P}$. On this space one can define a prior model of the unknowns and an observational model of the data that represent the encountered parameter uncertainties and the experimental situation, respectively.

The epistemic uncertainty of the parameter values is cast as a $\mathcal{D}_{\boldsymbol{x}}$-valued random vector $\boldsymbol{X} \colon \Omega \to \mathcal{D}_{\boldsymbol{x}} \subset \mathbb{R}^M$. Here, the components of $\boldsymbol{X} = (X_1, \ldots, X_M)^\top$ are $\mathcal{D}_{x_i}$-valued random variables $X_i \colon \Omega \to \mathcal{D}_{x_i} \subset \mathbb{R}$ for $i = 1, \ldots, M$. Since the data have not been processed at this stage, the joint density of $\boldsymbol{X} \sim \pi(\boldsymbol{x})$ is called the *prior density*. Similarly, a $\mathbb{R}^N$-valued random vector $\boldsymbol{Y} \colon \Omega \to \mathbb{R}^N$ represents the observables. The components of $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^\top$ are real-valued random variables $Y_i \colon \Omega \to \mathbb{R}$ for $i = 1, \ldots, N$. In order to draw inferences from the data about the unknowns, one has to formulate an observational model that establishes a relationship between those quantities. Commonly this is a probabilistic representation of the observables $\boldsymbol{Y}|\boldsymbol{x} \sim f(\boldsymbol{y}|\boldsymbol{x})$ that is conditional on the unknown parameters. For the actually acquired data $\boldsymbol{Y} = \boldsymbol{y}$, the *likelihood function* $\mathcal{L}(\boldsymbol{x}) = f(\boldsymbol{y}|\boldsymbol{x})$ is defined by interpreting the conditional density $f(\boldsymbol{y}|\boldsymbol{x})$ as a function of the unknowns $\boldsymbol{x}$.

Given this setup, one can formulate an updated probability density $\pi(\boldsymbol{x}|\boldsymbol{y})$ of the unknowns that is conditioned on the realized data. This so-called *posterior density* results from *Bayes' law*

$$\pi(\boldsymbol{x}|\boldsymbol{y}) = \frac{\mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})}{Z}. \tag{8.1}$$

It completely summarizes the available information about the unknowns after the data have been analyzed. The *model evidence* Z properly normalizes the posterior density. It can be written as

$$Z = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{L}(\boldsymbol{x})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{8.2}$$

One is often interested in the marginals and moments of the posterior. The posterior marginal $\pi(x_j|\boldsymbol{y})$ of a single unknown x_j with $j \in \{1, \ldots, M\}$ is defined as

$$\pi(x_j|\boldsymbol{y}) = \int_{\mathcal{D}_{\boldsymbol{x}_{\sim j}}} \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}_{\sim j}. \tag{8.3}$$

Here, the simplifying notation $\boldsymbol{x}_{\sim j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_M)^\top$ is introduced. For the mean $\mathbb{E}[\boldsymbol{X}|\boldsymbol{y}]$ and the covariance matrix $\mathrm{Cov}[\boldsymbol{X}|\boldsymbol{y}]$ of the posterior one has

$$\mathbb{E}[\boldsymbol{X}|\boldsymbol{y}] = \int_{\mathcal{D}_{\boldsymbol{x}}} \boldsymbol{x}\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}, \tag{8.4}$$

$$\mathrm{Cov}[\boldsymbol{X}|\boldsymbol{y}] = \int_{\mathcal{D}_{\boldsymbol{x}}} (\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{y}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{y}])^\top\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}. \tag{8.5}$$

More generally, Bayesian inference focuses the computation of posterior expectation values of the *quantities of interest* (QoI) $h \colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R}$. These expectations may be formally expressed as

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \int_{\mathcal{D}_{\boldsymbol{x}}} h(\boldsymbol{x})\, \pi(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}. \tag{8.6}$$

For later considerations, it is remarked that this integration over the posterior density can be interpreted as a reweighted integration over the prior density

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \int_{\mathcal{D}_{\boldsymbol{x}}} h(\boldsymbol{x}) \frac{\mathcal{L}(\boldsymbol{x})}{Z} \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \frac{1}{Z}\mathbb{E}[h(\boldsymbol{X})\mathcal{L}(\boldsymbol{X})]. \tag{8.7}$$

8.2.1 Bayesian inverse problems

The Bayesian framework described above can be applied to a vast range of scenarios from classical statistics [49, 50] and inverse modeling [51, 52]. In inverse problems, a so-called *forward model* $\mathcal{M}$ establishes a mathematical representation of the physical system under consideration. It is the function

$$\begin{aligned} \mathcal{M}\colon \mathcal{D}_{\boldsymbol{x}} &\to \mathbb{R}^N \\ \boldsymbol{x} &\mapsto \mathcal{M}(\boldsymbol{x}) \end{aligned} \tag{8.8}$$

that maps model inputs $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}} \subset \mathbb{R}^M$ to outputs $\tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{x}) \in \mathbb{R}^N$. *Inversion* is the process of inferring the unknown forward model parameters $\boldsymbol{x}$ with the measured data $\boldsymbol{y}$ of its response.

A probabilistic model of the observables is commonly constructed supposing that they can be represented as the sum $\boldsymbol{Y} = \tilde{\boldsymbol{Y}} + \boldsymbol{E}$ of the model response vector $\tilde{\boldsymbol{Y}} = \mathcal{M}(\boldsymbol{X})\colon \Omega \to \mathbb{R}^N$ and another random vector $\boldsymbol{E}\colon \Omega \to \mathbb{R}^N$. The latter accounts for measurement noise and forward model inadequacy. It is assumed that the *residual vector* $\boldsymbol{E}$ is statistically independent from $\boldsymbol{X}$. An unknown realization $\boldsymbol{E} = \boldsymbol{\varepsilon}$ measures the discrepancy between the actually measured data $\boldsymbol{y} = \tilde{\boldsymbol{y}} + \boldsymbol{\varepsilon}$ and the model response $\tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{x})$ at the true value $\boldsymbol{x}$. Typically one starts from the premise that the residual $\boldsymbol{E} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \boldsymbol{\Sigma})$ follows a Gaussian distribution. Here, $\boldsymbol{\Sigma}$ is a symmetric and positive-definite covariance matrix. The observational model is then simply given as $\boldsymbol{Y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}|\mathcal{M}(\boldsymbol{x}), \boldsymbol{\Sigma})$. For the likelihood this implies

$$\mathcal{L}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \mathcal{M}(\boldsymbol{x})\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y} - \mathcal{M}(\boldsymbol{x})\right)\right). \tag{8.9}$$

For the actually acquired data $\boldsymbol{y}$, this is understood as a function of the unknowns $\boldsymbol{x}$. The Bayesian solution to the inverse problem posed is then the posterior in Eq. (8.1) where the likelihood is given as in Eq. (8.9). It summarizes the collected information about the unknown forward model inputs.

8.2.2 Markov chain Monte Carlo

Apart from some exceptional cases, the posterior density in Eq. (8.1) does not exhibit a closed-form expression. Thus one settles either for computing expectation values under the posterior or for sampling from the posterior. The former can be accomplished through stochastic integration techniques such as *Monte Carlo* (MC) [53] or *importance sampling* [54]. For the latter one usually has to resort to *Markov chain Monte Carlo* (MCMC) sampling [9, 10]. An ergodic Markov chain $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots$ over the support $\mathcal{D}_{\boldsymbol{x}}$ is constructed in such a way that the posterior arises as the invariant distribution

$$\pi(\boldsymbol{x}^{(t+1)}|\boldsymbol{y}) = \int_{\mathcal{D}_{\boldsymbol{x}}} \pi(\boldsymbol{x}^{(t)}|\boldsymbol{y})\, \mathcal{K}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)})\, \mathrm{d}\boldsymbol{x}^{(t)}. \tag{8.10}$$

Here, $\mathcal{K}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)})$ denotes the density of the transition probability from the state $\boldsymbol{x}^{(t)}$ of the Markov chain at a time t to its state $\boldsymbol{x}^{(t+1)}$ at time $t+1$. The *Metropolis-Hastings* (MH) algorithm [55, 56] suggests an easy principle for the construction of a Markov kernel $\mathcal{K}$ that satisfies Eq. (8.10). It is based on sampling candidates from a proposal distribution and a subsequent accept/reject decision. The transition kernel defined this ways satisfies detailed balance, i.e. time reversibility $\pi(\boldsymbol{x}^{(t)}|\boldsymbol{y})\, \mathcal{K}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)}) = \pi(\boldsymbol{x}^{(t+1)}|\boldsymbol{y})\, \mathcal{K}(\boldsymbol{x}^{(t+1)}, \boldsymbol{x}^{(t)})$. This is a sufficient condition for Eq. (8.10) to apply. In practice, one initializes the Markov chain at some $\boldsymbol{x}^{(1)} \in \mathcal{D}_{\boldsymbol{x}}$ and then iteratively applies the MH updates from $\boldsymbol{x}^{(t)}$ to $\boldsymbol{x}^{(t+1)}$ for a finite number of times T. The *ergodic theorem* then ensures that one can approximate the population average in Eq. (8.6) in an asymptotically consistent way as the time average

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] \approx \frac{1}{T} \sum_{t=1}^{T} h(\boldsymbol{x}^{(t)}). \tag{8.11}$$

A whole string a unpleasant consequences is entailed by the fact that MCMC updates are typically local and serially dependent. The quality of the posterior approximation is governed by the MCMC sample autocorrelation. In order to ensure an efficient posterior exploration one has to carefully design and tune the proposal distribution. This is an extremely tedious and problem-dependent task. Yet, even for comparably efficient MCMC updates, a large number of MCMC iterations may be required in order to achieve an acceptable degree of fidelity of the final results. In inverse modeling this requires an even larger number of serial forward solves which can be prohibitively expensive for demanding models. Another intrinsic MCMC weakness is that it lacks a clear convergence and stopping criterion, i.e. for diagnosing when the chain has forgotten its initialization and has

converged to the target distribution in Eq. (8.10), and for the assessment of when the MC error in Eq. (8.11) has become sufficiently small. Even though there are more or less sophisticated convergence diagnostics [57, 58], those heuristic checks may very well fail, e.g. when separated posterior modes have not yet been detected.

The model evidence in Eq. (8.2) is important in the context of model comparison and selection [59]. In engineering applications it often happens that one wants to judge the performance of various competing models against measured data [60, 61]. While in variational Bayesian inference at least a lower bound of the model evidence is implicitly computed as a side product, in the MH algorithm it is not computed at all. Avoiding the explicit computation of the model evidence is beneficial for parameter estimation, but it does not allow for model selection. To this effect one has to rely on dedicated methods [62, 63].

8.3 Surrogate forward modeling

In the analysis of engineering systems it has become a widespread practice to substitute expensive computer models with inexpensive *metamodels* or *surrogate models*. Those approximations mimic the functional relationship between the inputs and the outputs of the original model in Eq. (8.8). Metamodeling promises significant gains in situations that require a large number of forward model runs, e.g. for optimization problems, uncertainty analysis and inverse modeling. Important classes of metamodels are based on Gaussian process models or Kriging [64, 65] and polynomial chaos expansions [66]. More recent introductions to these subjects can be found in [67, 68] and [69, 70], respectively. Nowadays the application of Kriging [16, 17] and polynomial chaos surrogates [18–20] is commonplace in Bayesian inverse problems.

We focus on polynomial chaos metamodels next. The idea is to decompose the forward model response into polynomial terms that are orthogonal with respect to a weight function. In stochastic analysis this weight is often identified with a probability density in order to facilitate uncertainty propagation. In inverse analysis it is commonly equated with the prior in order to enhance MCMC posterior sampling. The formalism of polynomial chaos expansions is rooted in spectral methods and functional approximations with orthogonal polynomials. Hence, the function space point of view is emphasized in this section. We also concentrate on linear least squares for the practical computation of the expansions coefficients.

8.3.1 L^2_π function space

From here on it is assumed that the components of the uncertain parameter vector $\boldsymbol{X} = (X_1, \dots, X_M)^\top$ are independent random variables X_i. Thus their joint density can be written as

$$\pi(\boldsymbol{x}) = \prod_{i=1}^{M} \pi_i(x_i). \tag{8.12}$$

Let $L^2_\pi(\mathcal{D}_{\boldsymbol{x}}) = \{u\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R} |\, \int_{\mathcal{D}_{\boldsymbol{x}}} u^2(\boldsymbol{x})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} < \infty\}$ be the Hilbert space of functions that are square integrable with respect to the prior density in Eq. (8.12). For $u, v \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ a weighted inner product $\langle \cdot, \cdot \rangle_{L^2_\pi}$ and its associated norm $\|\cdot\|_{L^2_\pi}$ are defined as

$$\langle u, v \rangle_{L^2_\pi} = \int_{\mathcal{D}_{\boldsymbol{x}}} u(\boldsymbol{x}) v(\boldsymbol{x})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{8.13}$$

$$\|u\|_{L^2_\pi} = \langle u, u \rangle^{1/2}_{L^2_\pi}. \tag{8.14}$$

Given that $u, v \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$, the real-valued random variables $u(\boldsymbol{X}), v(\boldsymbol{X})\colon \Omega \to \mathbb{R}$ on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$ have a finite variance. One can then write the inner product in Eq. (8.13) as the expectation

$$\langle u, v \rangle_{L^2_\pi} = \mathbb{E}[u(\boldsymbol{X}) v(\boldsymbol{X})]. \tag{8.15}$$

In the further course of the presentation, the identity in Eq. (8.15) is frequently used in order to switch back and forth between expectation values under the prior distribution and weighted inner products.

8.3.2 Orthonormal polynomials

Now a basis of the space $L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ is constructed with orthogonal polynomials [71–73]. Let $\{\Psi^{(i)}_{\alpha_i}\}_{\alpha_i \in \mathbb{N}}$ be a family of univariate polynomials in the input variable $x_i \in \mathcal{D}_{x_i}$. Each member is characterized by its polynomial

degree $\alpha_i \in \mathbb{N}$. The polynomials are required to be orthonormal in the sense that

$$\left\langle \Psi_{\alpha_i}^{(i)}, \Psi_{\beta_i}^{(i)} \right\rangle_{L^2_{\pi_i}} = \delta_{\alpha_i \beta_i} = \begin{cases} 1 & \text{if } \alpha_i = \beta_i, \\ 0 & \text{if } \alpha_i \neq \beta_i. \end{cases} \tag{8.16}$$

These polynomials form a complete orthogonal system in $L^2_{\pi_i}(\mathcal{D}_{x_i})$. Next, a set of multivariate polynomials $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^M}$ in the input variables $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ is constructed as the tensor product

$$\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \prod_{i=1}^{M} \Psi_{\alpha_i}^{(i)}(x_i). \tag{8.17}$$

Here, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M) \in \mathbb{N}^M$ is a multi-index that characterizes the polynomials. By construction, namely due to Eqs. (8.12), (8.16) and (8.17), they are orthonormal in the sense that

$$\langle \Psi_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\beta}} \rangle_{L^2_\pi} = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \begin{cases} 1 & \text{if } \boldsymbol{\alpha} = \boldsymbol{\beta}, \\ 0 & \text{if } \boldsymbol{\alpha} \neq \boldsymbol{\beta}. \end{cases} \tag{8.18}$$

These polynomials establish a complete orthogonal basis in $L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$. Note that the constant term is always given as $\Psi_0^{(i)} = 1$ in the univariate case. This ensures the proper normalization $\|\Psi_0^{(i)}\|_{L^2_{\pi_i}} = 1$. In the multivariate case one similarly has $\Psi_{\mathbf{0}} = 1$ with $\|\Psi_{\mathbf{0}}\|_{L^2_\pi} = 1$.

8.3.3 Hermite and Legendre polynomials

Two classical univariate families are the *Hermite* and the *Legendre polynomials* $\{H_{\alpha_i}(x_i)\}_{\alpha_i \in \mathbb{N}}$ for $x_i \in \mathbb{R}$ and $\{P_{\alpha_i}(x_i)\}_{\alpha_i \in \mathbb{N}}$ for $x_i \in [-1, 1]$, respectively. The former are orthogonal with respect to the weight function $\mathcal{N}(x_i|0,1) = (2\pi)^{-1/2} \exp(-x_i^2/2)$, the latter with respect to $\mathcal{U}(x_i|-1,1) = \mathrm{I}_{[-1,1]}(x_i)/2$. Here, $\mathrm{I}_{[-1,1]}$ denotes the indicator function of the interval $[-1, 1]$. A short summary of these two univariate families is given in Table 8.1. Over the respective domains, their first members are defined as given in Appendix 8.A. Classical orthogonal polynomials $\{\psi_{\alpha_i}^{(i)}(x_i)\}_{\alpha_i \in \mathbb{N}}$ are typically not normalized, e.g. the aforementioned Hermite or Legendre families. An orthonormal family $\{\Psi_{\alpha_i}^{(i)}\}_{\alpha_i \in \mathbb{N}}$ is then obtained through an appropriate normalization with $\Psi_{\alpha_i}^{(i)} = \psi_{\alpha_i}^{(i)} / \|\psi_{\alpha_i}^{(i)}\|_{L^2_{\pi_i}}$.

Table 8.1: Two families of orthogonal polynomials.

Input type	Polynomials	$\mathcal{D}_{x_i}$	$\pi_i(x_i)$	$\psi_{\alpha_i}^{(i)}(x_i)$	$\|\psi_{\alpha_i}^{(i)}\|_{L^2_{\pi_i}}$
Gaussian	Hermite	$\mathbb{R}$	$\mathcal{N}(x_i\|0,1)$	$H_{\alpha_i}(x_i)$	$\sqrt{\alpha_i!}$
Uniform	Legendre	$[-1,1]$	$\mathcal{U}(x_i\|-1,1)$	$P_{\alpha_i}(x_i)$	$\sqrt{1/(2\alpha_i+1)}$

In practice, the parameter space $\mathcal{D}_{\boldsymbol{x}}$ and the input distribution $\pi(\boldsymbol{x})$ are often not directly suitable for an expansion based on the two standardized families in Table 8.1. One possibility is then to employ suitably chosen or constructed polynomials [74, 75]. Another possibility is to use an invertible function $\mathcal{T}\colon \mathbb{R}^M \to \mathbb{R}^M$, sufficiently well-behaved and as linear as possible, in order to transform the physical variables $\boldsymbol{x}$ into *standardized variables* $\boldsymbol{\xi} = \mathcal{T}(\boldsymbol{x})$, i.e. the image $\mathcal{D}_{\boldsymbol{\xi}} = \mathcal{T}(\mathcal{D}_{\boldsymbol{x}})$ and the transformed weight function $\pi_{\mathcal{T}}(\boldsymbol{\xi}) = \pi(\mathcal{T}^{-1}(\boldsymbol{\xi}))\,|\det J_{\mathcal{T}^{-1}}(\boldsymbol{\xi})|$ are of a standard form. Here, $J_{\mathcal{T}^{-1}} = \mathrm{d}\mathcal{T}^{-1}/\mathrm{d}\boldsymbol{\xi}$ is the Jacobian matrix. If such a change of variables is needed, the considerations that are given below for $\mathcal{M}$ and h in the variables $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ can be straightforwardly repeated for $\mathcal{M} \circ \mathcal{T}^{-1}$ and $h \circ \mathcal{T}^{-1}$ in the variables $\boldsymbol{\xi} \in \mathcal{D}_{\boldsymbol{\xi}}$. In this case, the expectation in Eq. (8.6) follows the integration by substitution

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \frac{1}{Z} \int_{\mathcal{D}_{\boldsymbol{\xi}}} h\left(\mathcal{T}^{-1}(\boldsymbol{\xi})\right) \mathcal{L}\left(\mathcal{T}^{-1}(\boldsymbol{\xi})\right)\, \pi_{\mathcal{T}}(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi}. \tag{8.19}$$

8.3.4 Polynomial chaos expansions

For simplicity, herein the presentation is restricted to scalar-valued models $\mathcal{M}\colon \mathcal{D}_{\boldsymbol{x}} \to \mathbb{R}$. The extension to the multivariate case is straightforward. It is supposed that the forward model $\mathcal{M} \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ is mean-square

integrable. This is a reasonable assumption as seen from a physical perspective. In $L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ the *generalized Fourier expansion* of $\mathcal{M}$ in terms of the orthogonal polynomials $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^M}$ is then given as

$$\mathcal{M} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} a_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}, \quad \text{with} \tag{8.20}$$

$$a_{\boldsymbol{\alpha}} = \langle \mathcal{M}, \Psi_{\boldsymbol{\alpha}}\rangle_{L^2_\pi} = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{M}(\boldsymbol{x})\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{8.21}$$

The *generalized Fourier coefficients* $\{a_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^M}$ of the series expansion in Eq. (8.20) are defined as the orthogonal projection of $\mathcal{M}$ onto the basis elements in Eq. (8.21). The corresponding Fourier series of the second-order random variable $\tilde{Y} = \mathcal{M}(\boldsymbol{X})$ on $(\Omega, \mathcal{F}, \mathcal{P})$ is a so-called *polynomial chaos expansion* (PCE) $\mathcal{M}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} a_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{X})$.

PCEs have been popularized in the context of uncertainty propagation where the goal is the quantification of the distribution of $\tilde{Y} = \mathcal{M}(\boldsymbol{X})$. For this purpose it comes in handy that the mean and the variance of this random variable can be easily determined from its PCE coefficients. Indeed, with Eq. (8.18) it is easy to verify that they are simply given as

$$\mathbb{E}[\mathcal{M}(\boldsymbol{X})] = \left\langle \Psi_{\boldsymbol{0}}, \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} a_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}} \right\rangle_{L^2_\pi} = a_{\boldsymbol{0}}, \tag{8.22}$$

$$\mathrm{Var}[\mathcal{M}(\boldsymbol{X})] = \left\| \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} a_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}} - a_{\boldsymbol{0}}\Psi_{\boldsymbol{0}} \right\|^2_{L^2_\pi} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M\setminus\{\boldsymbol{0}\}} a^2_{\boldsymbol{\alpha}}. \tag{8.23}$$

The simple identities in Eqs. (8.22) and (8.23) follow from the definitions of the inner product and the associated norm in Eqs. (8.13) and (8.14), respectively.

8.3.5 Truncated series

For a practical computation one has to truncate the infinite series in Eq. (8.20). Let the total degree of a multivariate polynomial $\Psi_{\boldsymbol{\alpha}}$ be defined as $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^M |\alpha_i|$. A *standard truncation scheme* is then adopted by limiting the terms in Eq. (8.20) to the finite set of multi-indices

$$\mathcal{A}_p = \left\{\boldsymbol{\alpha}\in\mathbb{N}^M : \|\boldsymbol{\alpha}\|_1 \le p\right\}. \tag{8.24}$$

This specifies a set of polynomials $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}_p}$ such that their total degree $\|\boldsymbol{\alpha}\|_1$ is smaller than or equal to a chosen p. The total number of terms retained in the set $\mathcal{A}_p$ is given as

$$P = \binom{M+p}{p} = \frac{(M+p)!}{M!\,p!}. \tag{8.25}$$

The dramatic increase of the total number of terms P with the input dimensionality M and the maximal polynomial degree p, that is described by Eq. (8.25), is commonly referred to as the *curse of dimensionality*. A simple idea to limit the number of regressors relies on hyperbolic truncation sets. For $0 < q < 1$ a quasinorm is defined as $\|\boldsymbol{\alpha}\|_q = (\sum_{i=1}^M |\alpha_i^q|)^{1/q}$. The corresponding *hyperbolic truncation scheme* is then given as $\mathcal{A}_p^q = \{\boldsymbol{\alpha}\in\mathbb{N}^M \,|\, \|\boldsymbol{\alpha}\|_q \le p\}$. Adopting the standard scheme in Eq. (8.24), a finite version of Eq. (8.20) can be written as

$$\hat{\mathcal{M}}_p(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}_p} a_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}). \tag{8.26}$$

In engineering problems, one uses $\hat{\mathcal{M}}_p(\boldsymbol{x})$ as a functional approximation of $\mathcal{M}(\boldsymbol{x})$, i.e. as a *polynomial response surface* [76]. This is justified because the approximation converges in the mean-square sense

$$\left\|\mathcal{M} - \hat{\mathcal{M}}_p\right\|^2_{L^2_\pi} = \mathbb{E}\left[\left(\mathcal{M}(\boldsymbol{X}) - \hat{\mathcal{M}}_p(\boldsymbol{X})\right)^2\right] = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M\setminus\mathcal{A}_p} a^2_{\boldsymbol{\alpha}} \to 0, \quad \text{for } p\to\infty. \tag{8.27}$$

The rate of the convergence in Eq. (8.27) depends on the regularity of $\mathcal{M}$. On top of that, the response surface in Eq. (8.26) is also optimal in the mean-square sense

$$\left\|\mathcal{M} - \hat{\mathcal{M}}_p\right\|^2_{L^2_\pi} = \inf_{\mathcal{M}^\star\in\mathbb{P}_p} \|\mathcal{M} - \mathcal{M}^\star\|^2_{L^2_\pi}, \quad \text{where } \mathbb{P}_p = \mathrm{span}\left(\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}_p}\right). \tag{8.28}$$

According to Eq. (8.28), the response surface in Eq. (8.26) minimizes the mean-square error over the space of polynomials $\mathbb{P}_p = \mathrm{span}(\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}_p})$ having a total degree of at most p.

8.3.6 Least squares

In order to find a metamodel of the form Eq. (8.26), one computes approximations of the exact expansion coefficients in Eq. (8.21). Broadly speaking, one distinguishes between *intrusive* and *non-intrusive* computations. While the former class of techniques is based on manipulations of the governing equations, the latter is exclusively build upon calls to the forward model at chosen input values. Stochastic Galerkin methods belong to the class of intrusive techniques [77, 78], whereas stochastic collocation [79, 80] and projection through numerical quadrature [81, 82] are non-intrusive approaches. Herein we focus on another non-intrusive formulation that is based on least squares regression analysis [83, 84]. This formulation is based on linear least squares [37, 38] and related ideas from statistical learning theory [39, 40]. Since this includes sparsity-promoting fitting techniques from high-dimensional statistics [85, 86], recently least squares projection methods receive considerable attention. This includes frequentist [87–90] and Bayesian implementations [91–94] of shrinkage estimators. Current results regarding the convergence behavior of such regression methods can be found in [95–97].

We introduce a simplifying vector notation such that $\boldsymbol{a} = (a_1, \dots, a_P)^\top$ and $\boldsymbol{\Psi} = (\Psi_1, \dots, \Psi_P)^\top$ gather and order the coefficients and the polynomials for all $\boldsymbol{\alpha} \in \mathcal{A}_p$. For the truncated expression in Eq. (8.26) one thus has $\hat{\mathcal{M}}_p = \boldsymbol{a}^\top \boldsymbol{\Psi}$. The problem of finding $\hat{\mathcal{M}}_p \in \mathbb{P}_p$ that minimizes the mean-square error in Eq. (8.28) may then be equivalently rephrased as

$$\boldsymbol{a} = \underset{\boldsymbol{a}^\star \in \mathbb{R}^P}{\arg\min}\, \mathbb{E}\left[\left(\mathcal{M}(\boldsymbol{X}) - \boldsymbol{a}^{\star\top} \boldsymbol{\Psi}(\boldsymbol{X})\right)^2\right]. \tag{8.29}$$

The stochastic optimization objective in Eq. (8.29) establishes an alternative to the orthogonal projection in Eq. (8.21). This formulation may be more amenable to a numerical computation. At the very least it allows one the draw on the machinery of linear least squares in order to compute an approximation $\hat{\boldsymbol{a}}$ of the exact coefficients $\boldsymbol{a}$. To that end one discretizes Eq. (8.29) as

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}^\star \in \mathbb{R}^P}{\arg\min}\, \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{M}(\boldsymbol{x}^{(k)}) - \boldsymbol{a}^{\star\top} \boldsymbol{\Psi}(\boldsymbol{x}^{(k)})\right)^2. \tag{8.30}$$

Here, the *experimental design* $\mathcal{X} = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(K)})$ is a representative sample of K forward model inputs, e.g. randomly drawn from the input distribution in Eq. (8.12). It is then required to compute the corresponding model responses $\mathcal{Y} = (\mathcal{M}(\boldsymbol{x}^{(1)}), \dots, \mathcal{M}(\boldsymbol{x}^{(K)}))^\top$ in K training runs.

Now let $\boldsymbol{A} \in \mathbb{R}^{K \times P}$ be the matrix with entries $A_{k,l} = \Psi_l(\boldsymbol{x}^{(k)})$ for $k = 1, \dots, K$ and $l = 1, \dots, P$. Moreover, let the system $\mathcal{Y} = \boldsymbol{A}\hat{\boldsymbol{a}}$ be overdetermined with $K \geq P$. The linear least squares problem in Eq. (8.30) may then be written as

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}^\star \in \mathbb{R}^P}{\arg\min} \|\mathcal{Y} - \boldsymbol{A}\boldsymbol{a}^\star\|^2 \tag{8.31}$$

The *normal equations* $(\boldsymbol{A}^\top \boldsymbol{A})\hat{\boldsymbol{a}} = \boldsymbol{A}^\top \mathcal{Y}$ establish the first-order condition for Eq. (8.31) to apply. Given that the matrix $\boldsymbol{A}^\top \boldsymbol{A}$ is non-singular, this linear system is solved by

$$\hat{\boldsymbol{a}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \mathcal{Y}. \tag{8.32}$$

The positive definiteness of $\boldsymbol{A}^\top \boldsymbol{A}$, i.e. the columns of $\boldsymbol{A}$ are linearly independent, is the second-order condition for Eq. (8.32) to be the minimum of Eq. (8.31). The *ordinary least squares* (OLS) solution Eq. (8.32) is commonly computed by means of linear algebra methods. An alternative to OLS is *least angle regression* (LAR) [98, 99]. LAR is well-suited for high-dimensional PCE regression problems [87] and can be even applied in the underdetermined case. It is based on selecting only the most dominant regressors from a possibly large candidate set. The resulting predictor is thus sparse as compared to the OLS solution.

8.3.7 Prediction errors

After the computation of a metamodel, one typically wants to assess its prediction accuracy. Moreover, when a number of candidate surrogates is computed, one wants to compare their performances in order to eventually select the best. Hence, one needs to define an appropriate criterion that allows for an accurate and efficient quantification of the approximation errors. The natural measure of the mismatch between the forward model $\mathcal{M}$ and an approximation $\hat{\mathcal{M}}_p$ is the *generalization error* $E_{\text{Gen}} = \mathbb{E}[(\mathcal{M}(\boldsymbol{X}) - \hat{\mathcal{M}}_p(\boldsymbol{X}))^2]$. This is exactly the error the minimization of which is posed by Eq. (8.29). Since it remains unknown, it cannot be used as a performance measure. One could estimate E_{Gen} based on MC simulation, though. However, this is not very efficient since it requires the execution of additional forward model runs. In contrast, the *empirical error* $E_{\text{Emp}} = K^{-1} \sum_{k=1}^{K} (\mathcal{M}(\boldsymbol{x}^{(k)}) - \hat{\mathcal{M}}_p(\boldsymbol{x}^{(k)}))^2$ is the quantity that is practically minimized according to Eq. (8.30).

This error indicator is obtained for free, however, it does not account for overfitting and thus tends to severely underestimate the real generalization error E_{Gen}.

In order to construct an estimate of E_{Gen} that is more efficient than the MC estimate and more accurate than E_{Emp}, one sometimes resorts to leave-one-out (LOO) cross validation [100]. Let $\hat{\mathcal{M}}_{\sim k}$ be the surrogate model that is obtained from the reduced experimental design $\mathcal{X}_{\sim k} = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(k-1)}, \boldsymbol{x}^{(k+1)}, \dots, \boldsymbol{x}^{(K)})$, i.e. a single input $\boldsymbol{x}^{(k)}$ has been dropped. The *LOO error* is then defined as

$$E_{\text{LOO}} = \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{M}(\boldsymbol{x}^{(k)}) - \hat{\mathcal{M}}_{\sim k}(\boldsymbol{x}^{(k)}) \right)^2 . \tag{8.33}$$

Without the need for re-running the forward model, this error allows for a fair assessment of how well the performance of a metamodel $\hat{\mathcal{M}}_p$ generalizes beyond the used experimental design. Yet, Eq. (8.33) calls for conducting K separate regressions for finding $\hat{\mathcal{M}}_{\sim k}$ with an experimental design of the size $K-1$. A remarkably simple result from linear regression analysis states that E_{LOO} can be also computed as

$$E_{\text{LOO}} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\mathcal{M}(\boldsymbol{x}^{(k)}) - \hat{\mathcal{M}}_p(\boldsymbol{x}^{(k)})}{1 - h_k} \right)^2 . \tag{8.34}$$

Here, $\hat{\mathcal{M}}_p$ is computed from the full experimental design $\mathcal{X}$ and h_k denotes the k-th diagonal entry of the matrix $\boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top$. This is more efficient since it does not require repeated fits. A derivation of the formula in Eq. (8.34) can be found in [101].

One may define $\epsilon_{\text{Emp}} = E_{\text{Emp}}/\text{Var}[\mathcal{Y}]$ and $\epsilon_{\text{LOO}} = E_{\text{LOO}}/\text{Var}[\mathcal{Y}]$ as normalized versions of the empirical and the LOO error, respectively. Here, $\text{Var}[\mathcal{Y}]$ is the empirical variance of the response sample $\mathcal{Y}$. These normalized errors can be used in order to judge and compare the performance of metamodels. In turn, this enables a practical convergence analysis, e.g. by monitoring the errors over repeated metamodel computations for an increasing experimental design size K and expansion order p.

8.4 Spectral Bayesian inference

Different types of probability density approximations are encountered in statistical inference. This includes series expansions for the population density of a random data sample in nonparametric distribution fitting. Here, the unknown density of the data is either directly represented as a linear combination of polynomial basis functions [102] or as the product of a base density times a superposition of polynomials [103]. The latter type of expansion is also encountered in parametric density estimation of random data where one-dimensional posterior densities of the unknown parameter are expanded about a Gaussian baseline. This can be based on a Taylor sum at a maximum likelihood estimate [104, 105] or on Stein's lemma and numerical integration [106, 107]. Moreover, different types of likelihood approximations are encountered in inverse modeling. This includes direct approaches where the likelihood is approximated itself [108, 109] and indirect methods where the likelihood is approximated based on a surrogate of the forward model [18–20]. These techniques facilitate Bayesian inference within the limits of MCMC sampling.

By linking likelihood approximations to density expansions, now we present a spectral formulation of Bayesian inference which targets the emulation of the posterior density. Based on the theoretical and computational machinery of PCEs, the likelihood function itself is decomposed into polynomials that are orthogonal with respect to the prior distribution. This spectral likelihood expansion enables semi-analytic Bayesian inference. Simple formulas are derived for the joint posterior density and its marginals. They are regarded as expansions of the posterior about the prior as the reference density. The model evidence is shown to be the coefficient of the constant expansion term. General QoI-expectations under the posterior and the first posterior moments are obtained through a mere postprocessing of the spectral coefficients. After a discussion of the advantages and shortcomings of the spectral method, a change of the reference density is proposed in order to improve the efficacy.

8.4.1 Spectral likelihood expansions

The authors C. Soize and R. Ghanem start their paper [110] with the following sentence: "Characterizing the membership of a mathematical function in the most suitable functional space is a critical step toward analyzing it and identifying sequences of efficient approximants to it." As discussed in Section 8.2, given the data $\boldsymbol{y}$ and

the statistical model $f(\boldsymbol{y}|\boldsymbol{x})$, the likelihood is the function

$$\begin{aligned} \mathcal{L}\colon \mathcal{D}_{\boldsymbol{x}} &\to \mathbb{R}^+ \\ \boldsymbol{x} &\mapsto f(\boldsymbol{y}|\boldsymbol{x}). \end{aligned} \tag{8.35}$$

It maps the parameter space $\mathcal{D}_{\boldsymbol{x}}$ into the set of non-negative real numbers $\mathbb{R}^+$. At this place we assume that the likelihood $\mathcal{L} \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ is square integrable with respect to the prior. From a statistical point of view, this is a reasonable supposition which is necessary in order to invoke the theory of Section 8.3. On condition that the likelihood is bounded from above, the mean-square integrability follows immediately from the axioms of probability. Note that maximum likelihood estimates (MLE) implicitly rest upon this presumption. If $\boldsymbol{x}_{\mathrm{MLE}} \in \arg\max_{\boldsymbol{x}\in\mathcal{D}_{\boldsymbol{x}}} \mathcal{L}(\boldsymbol{x})$ is a MLE, i.e. for all $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ it applies that $\mathcal{L}(\boldsymbol{x}) \leq \mathcal{L}(\boldsymbol{x}_{\mathrm{MLE}}) < \infty$, then one trivially has $\int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{L}^2(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \leq \mathcal{L}^2(\boldsymbol{x}_{\mathrm{MLE}}) \int_{\mathcal{D}_{\boldsymbol{x}}} \pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathcal{L}^2(\boldsymbol{x}_{\mathrm{MLE}}) < \infty$.

Having identified $L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ as a suitable function space for characterizing the likelihood, one can represent the likelihood with respect to the orthonormal basis $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^M}$. This representation is

$$\mathcal{L} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} b_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}, \quad \text{with} \tag{8.36}$$

$$b_{\boldsymbol{\alpha}} = \langle \mathcal{L}, \Psi_{\boldsymbol{\alpha}} \rangle_{L^2_\pi} = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{L}(\boldsymbol{x}) \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{8.37}$$

We refer to Eqs. (8.36) and (8.37) as a *spectral likelihood expansion* (SLE). Notice that the SLE coefficients $\{b_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^M}$ are data-dependent. This reflects the fact that the likelihood in Eq. (8.35) depends on the data. With the truncation scheme in Eq. (8.24), one can limit the infinite series in Eq. (8.36) to the finite number of terms for which $\boldsymbol{\alpha} \in \mathcal{A}_p$. A mean-square convergent response surface of the likelihood is then given as

$$\hat{\mathcal{L}}_p(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}_p} b_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}). \tag{8.38}$$

For the time being we assume that the coefficients of the SLE in Eq. (8.36) or its response surface in Eq. (8.38) are already known. One can then accomplish Bayesian inference by extracting the joint posterior density or a *posterior density surrogate*, its marginals and the corresponding QoI-expectations directly from the SLE.

8.4.2 Joint posterior density

We begin with the joint posterior density function and the model evidence. By plugging Eq. (8.36) in Eq. (8.1) one simply obtains the "nonparametric" expression

$$\pi(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z} \left(\sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} b_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) \right) \pi(\boldsymbol{x}). \tag{8.39}$$

Due to the orthonormality of the basis, the model evidence is simply found as the coefficient of the constant SLE term. This is easily verified by writing Eq. (8.2) as

$$Z = \langle 1, \mathcal{L} \rangle_{L^2_\pi} = \left\langle \Psi_{\boldsymbol{0}}, \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} b_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}} \right\rangle_{L^2_\pi} = b_{\boldsymbol{0}}. \tag{8.40}$$

The remarkably simple result in Eq. (8.40) completes the expression of the posterior density in Eq. (8.39). It is interesting to note that the posterior density is of the form

$$\pi(\boldsymbol{x}|\boldsymbol{y}) = \pi(\boldsymbol{x}) \left(1 + \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M\setminus\{\boldsymbol{0}\}} b_{\boldsymbol{0}}^{-1} b_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) \right). \tag{8.41}$$

In essence, the posterior density is here represented as a "perturbation" around the prior. The latter establishes the leading term of the expansion and acts as the *reference density*. The expression in Eq. (8.41) is reminiscent of an *Edgeworth expansion* for a density function in asymptotic statistics [111–113].

8.4.3 Quantities of interest

Based on the joint posterior density one can calculate the corresponding QoI-expectations in Eq. (8.6). At this point, the identities in Eqs. (8.7) and (8.15) finally play a key role. They allow one to express and treat the posterior expectation of a QoI with $h \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ as the weighted projection onto the likelihood

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \frac{1}{Z}\mathbb{E}[h(\boldsymbol{X})\mathcal{L}(\boldsymbol{X})] = \frac{1}{Z}\left\langle h, \mathcal{L}\right\rangle_{L^2_\pi}. \tag{8.42}$$

Let $h = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} c_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}$ with $c_{\boldsymbol{\alpha}} = \langle h, \Psi_{\boldsymbol{\alpha}}\rangle_{L^2_\pi}$ be the QoI representation in the polynomial basis used. The general posterior expectation in Eq. (8.6) follows then from Eq. (8.42) by *Parseval's theorem*

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \frac{1}{b_{\boldsymbol{0}}}\left\langle \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} c_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}, \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} b_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}\right\rangle_{L^2_\pi} = \frac{1}{b_{\boldsymbol{0}}}\sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} c_{\boldsymbol{\alpha}}b_{\boldsymbol{\alpha}}. \tag{8.43}$$

If the QoI $h \in \mathbb{P}_p$ is known by a multivariate monomial representation of finite order, then its representation in the orthogonal basis can always be recovered by a change of basis. Examples that relate to the first posterior moments are given shortly hereafter. In a more complex case the QoI is a computational model itself and one would have to numerically compute a PCE surrogate. While Eq. (8.22) facilitates the propagation of the prior uncertainty, Eq. (8.43) promotes the propagation of the posterior uncertainty.

8.4.4 Posterior marginals

Now the posterior marginals in Eq. (8.3) are derived. For some $j \in \{1,\dots,M\}$ let us introduce the new set of multi-indices $\mathcal{A}^{(j)} = \{(\alpha_1,\dots,\alpha_M)|\alpha_i = 0 \text{ for } i \neq j\}$. With a slight abuse of the notation, the *sub-expansion* of the SLE that only contains terms with $\boldsymbol{\alpha} \in \mathcal{A}^{(j)}$ is denoted as

$$\mathcal{L}_j(x_j) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}^{(j)}} b_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{\mu\in\mathbb{N}} b^{(j)}_\mu \Psi^{(j)}_\mu(x_j), \quad \text{where } b^{(j)}_\mu = b_{(0,\dots,0,\mu,0,\dots,0)}. \tag{8.44}$$

It collects all the polynomials that are constant in all variables $\boldsymbol{x}_{\sim j}$, i.e. they are non-constant in the single variable x_j only. In this sense the sub-expansion in Eq. (8.44) is virtually a function of x_j only. For the posterior marginal of the single unknown x_j in Eq. (8.3) one can derive

$$\pi(x_j|\boldsymbol{y}) = \frac{\pi_j(x_j)}{Z}\int_{\mathcal{D}_{\boldsymbol{x}_{\sim j}}} \mathcal{L}(\boldsymbol{x})\,\pi(\boldsymbol{x}_{\sim j})\,\mathrm{d}\boldsymbol{x}_{\sim j} = \frac{1}{b_{\boldsymbol{0}}}\mathcal{L}_j(x_j)\pi_j(x_j). \tag{8.45}$$

These equalities apply due to the independent prior $\pi(\boldsymbol{x}) = \pi(\boldsymbol{x}_{\sim j})\pi_j(x_j)$ in Eq. (8.12), the orthonormality of the univariate polynomials in Eq. (8.16) and the tensor structure of the multivariate ones in Eq. (8.17). For a pair $j,k \in \{1,\dots,M\}$ with $j \neq k$ let us introduce yet another set of multi-indices $\mathcal{A}^{(j,k)} = \{(\alpha_1,\dots,\alpha_M)|\alpha_i = 0 \text{ for } i \neq j,k\}$. The sub-expansion of the full SLE that only contains terms with $\boldsymbol{\alpha} \in \mathcal{A}^{(j,k)}$ is denoted as

$$\begin{gathered}\mathcal{L}_{j,k}(x_j,x_k) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}^{(j,k)}} b_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{\mu,\nu\in\mathbb{N}} b^{(j,k)}_{\mu,\nu}\Psi^{(j)}_\mu(x_j)\Psi^{(k)}_\nu(x_k),\\ \text{where } b^{(j,k)}_{\mu,\nu} = b_{(0,\dots,0,\mu,0,\dots,0,\nu,0,\dots,0)}.\end{gathered} \tag{8.46}$$

Since it only contains terms that are constant in $\boldsymbol{x}_{\sim j,k}$, the sub-expansion in Eq. (8.46) can be seen as a function of x_j and x_k. The posterior density can then be marginalized as follows

$$\pi(x_j,x_k|\boldsymbol{y}) = \int_{\mathcal{D}_{\boldsymbol{x}_{\sim j,k}}} \pi(\boldsymbol{x}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}_{\sim j,k} = \frac{1}{b_{\boldsymbol{0}}}\mathcal{L}_{j,k}(x_j,x_k)\pi_j(x_j)\pi_k(x_k). \tag{8.47}$$

Note that the dependency structure of $\pi(x_j,x_k|\boldsymbol{y})$ in Eq. (8.47) is induced by those terms of $\mathcal{L}_{j,k}(x_j,x_k)$ that are not present in $\mathcal{L}_j(x_j)$ and $\mathcal{L}_k(x_k)$, i.e. the terms $b^{(j,k)}_{\mu,\nu}\Psi^{(j)}_\mu(x_j)\Psi^{(k)}_\nu(x_k)$ with $\mu,\nu \neq 0$.

8.4.5 First posterior moments

With the two marginalizations of the posterior density in Eqs. (8.45) and (8.47) one can calculate the entries of the posterior mean in Eq. (8.4) and the covariance matrix in Eq. (8.5). Let $\{d^{(j)}_0, d^{(j)}_1\}$ be defined such that $x_j = d^{(j)}_0\Psi^{(j)}_0(x_j) + d^{(j)}_1\Psi^{(j)}_1(x_j)$. With this univariate representation and Eq. (8.45) one easily obtains

$$\mathbb{E}[X_j|\boldsymbol{y}] = \frac{1}{b_{\boldsymbol{0}}}\left\langle x_j, \mathcal{L}_j\right\rangle_{L^2_{\pi_j}} = \frac{1}{b_{\boldsymbol{0}}}\left(d^{(j)}_0 b^{(j)}_0 + d^{(j)}_1 b^{(j)}_1\right). \tag{8.48}$$

Note that one actually has $b_0^{(j)} = b_{\mathbf{0}}$ in this notation. Diagonal entries of the covariance matrix in Eq. (8.5) can be similarly deduced. Let $\{e_0^{(j)}, e_1^{(j)}, e_2^{(j)}\}$ the coefficients of the univariate representation $(x_j - \mathbb{E}[X_j|\boldsymbol{y}])^2 = \sum_{\mu=0}^{2} e_\mu^{(j)} \Psi_\mu^{(j)}(x_j)$. Then one simply has

$$\mathrm{Var}[X_j|\boldsymbol{y}] = \frac{1}{b_{\mathbf{0}}} \left\langle (x_j - \mathbb{E}[X_j|\boldsymbol{y}])^2, \mathcal{L}_j \right\rangle_{L^2_{\pi_j}} = \frac{1}{b_{\mathbf{0}}} \sum_{\mu=0}^{2} e_\mu^{(j)} b_\mu^{(j)}. \tag{8.49}$$

Finally, let $\{e_{0,0}^{(j,k)}, e_{0,1}^{(j,k)}, e_{1,0}^{(j,k)}, e_{1,1}^{(j,k)}\}$ be the coefficients of the bivariate PCE with $(x_j - \mathbb{E}[X_j|\boldsymbol{y}])(x_k - \mathbb{E}[X_k|\boldsymbol{y}]) = \sum_{\mu,\nu=0}^{1} e_{\mu,\nu}^{(j,k)} \Psi_\mu^{(j)}(x_j) \Psi_\nu^{(k)}(x_k)$. For an off-diagonal entry of Eq. (8.5) one then finds

$$\mathrm{Cov}[X_j, X_k|\boldsymbol{y}] = \frac{1}{b_{\mathbf{0}}} \sum_{\mu,\nu=0}^{1} e_{\mu,\nu}^{(j,k)} b_{\mu,\nu}^{(j,k)}. \tag{8.50}$$

Notation-wise, Eqs. (8.48) to (8.50) may seem to be somewhat cumbersome. Nevertheless, they establish simple recipes of how to obtain the first posterior moments by a postprocessing of the low-degree SLE terms in closed-form. Higher-order moments could be obtained similarly. Some examples of how the corresponding QoIs can be represented in terms of orthogonal polynomials can be found in Appendix 8.B.

8.4.6 Discussion of the advantages

In spectral Bayesian inference the posterior is genuinely characterized through the SLE and its coefficients. The essential advantage of this approach is that all quantities of inferential relevance can be computed semi-analytically. Simple formulas for the joint posterior density and the model evidence emerge in Eqs. (8.39) and (8.40). They allow to establish Eq. (8.41) as the posterior density surrogate. General QoI-expectations under the posterior are then calculated via Parseval's formula in Eq. (8.43). The posterior marginals are obtained based on sub-expansions of the full SLE in Eq. (8.45) and the first posterior moments have closed-form expressions in Eqs. (8.48) to (8.50). These striking characteristics clearly distinguish spectral inference from integration and sampling approaches where the posterior is summarized by expected values or random draws only. As for the latter, one has to rely on kernel estimates of the posterior density and on empirical sample approximations of the QoI-expectations. Also, the model evidence is not computed explicitly.

The practical computation of the SLE in Eq. (8.36) can be accomplished analogously to finding the PCE approximation of the forward model in Eq. (8.20), e.g. by solving a linear least squares problem as in Eq. (8.30). This allows one to draw on the vast number of tools that were developed for carrying out this well-known type of regression analysis. An attractive feature of this procedure is that the prediction error of the obtained SLE acts as a natural convergence indicator. We recall that the LOO error in Eq. (8.33) can be efficiently evaluated as per Eq. (8.34), i.e. without the need for additional forward model runs or regression analyses. The existence of an intrinsic convergence criterion is an advantage over traditional MCMC techniques. Another advantage of the formulation its amenability to parallel computations. While the workload posed by MCMC is inherently serial, running the forward model for each input in the experimental design is embarrassingly parallel. Parallelization is also possible on the level of the linear algebra operations that are necessary in order to solve the normal equations.

8.4.7 Discussion of the shortcomings

The approximate nature of SLE computations is twofold, i.e. only a finite number of terms are kept in the expansion and the coefficients are inexact. Unfortunately, a number of inconveniences may arise from these inevitable approximations. The SLE and the correspondingly computed posterior density could spuriously take on negative values. Also the estimated model evidence in Eq. (8.40) could take on negative values $Z < 0$. Still note that the approximate posterior in Eq. (8.39) always integrates to one. For reasonably adequate SLEs, we expect that negative values only occur in the distributional tails. Even so, the presence of negative values hampers the interpretation of the obtained posterior surrogate as a proper probability density, e.g. it leads to finite negative probabilities that are somehow irritating. From a more practical rather than a technical or philosophical perspective, densities are ultimately instrumental to the evaluation of more concrete quantities such as the expectations in Eq. (8.43). The severity of negative densities has thus to be judged with respect to the distortions of these relevant values. As long as their accurate approximation is guaranteed, the possibility of negative density values is an unavoidable artifact that can be regarded as a minor blemish. And the obtained surrogate density still proves to be expedient to effectively characterize the posterior distribution. In this light,

it is more unpleasant that the a posteriori estimates of the model parameters may violate the restrictions that were imposed a priori. In Eq. (8.48) it could indeed happen that $\mathbb{E}[X_j|\boldsymbol{y}] \notin \mathcal{D}_{x_j}$. Estimations of the second order moments in Eq. (8.49) could result in unnatural values $\mathrm{Var}[X_j|\boldsymbol{y}] < 0$, too. Although these problems cannot be remedied unless one solves an appropriately constrained version of Eq. (8.30), they unlikely occur if the SLE is sufficiently accurate. Anticipating outcomes from later numerical demonstrations, we remark that the occurrence of negative density values is observed, while unphysical or unnatural estimates of the first posterior moments are not found.

The SLE decomposition into a globally smooth basis of tensorized polynomials suffers from some other intrinsic problems. Generally, there is the curse of dimensionality, i.e. the increase of the number of regressors in Eq. (8.25). Furthermore, the SLE convergence rate in Eq. (8.27) depends on the regularity of the underlying likelihood function. For discontinuous forward or error models the SLE approximation with smooth polynomials converges only slowly. Likelihood functions often show a peaked structure around the posterior modes and a vanishing behavior elsewhere. Hence, any adequate superposition of polynomials has to capture those two different behavioral patterns through some kind of "constructive" and "destructive" interaction between its terms, respectively. Due to their global nature, the employed polynomial basis may not admit sparse likelihood representations. In turn, a high number of terms might be necessary in order to accurately represent even simple likelihood functions. Especially in the case of high-dimensional and unbounded parameter spaces this may cause severe practical problems. Of course, in Eqs. (8.36) and (8.37) one could expand the likelihood in a different basis. Yet, note that the QoIs in Eq. (8.43) would also have to be expanded in that basis.

The role of the prior for spectral inference is manifold. Initially the posterior expectations in Eq. (8.6) have been rewritten as the weighted prior expectations in Eq. (8.7). This formulation is accompanied by difficulties in computing posteriors that strongly deviate from the prior. The same situation arises for crude MC integration and eventually motivates importance or MCMC sampling that allow to focus on localized regions of high posterior mass. Those difficulties become manifest if the prior acts as the sampling distribution for the experimental design. In this case, the SLE is only accurate over the regions of the parameter space that accumulate considerable shares of the total prior probability mass. Approximation errors of the SLE in regions that are less supported by the prior then induce errors in the computed posterior surrogate. Note that this difficulty is also encountered in MCMC posterior exploration with prior-based PCE metamodels. Also, the error estimate in Eq. (8.34) then only measures the SLE accuracy with respect to the prior which may be misleading for the assessment of the posterior accuracy. It is not clear how the errors of the likelihood expansion relate to the induced errors of the posterior surrogate and the posterior moments. Moreover, since the prior acts as the reference density of the posterior expansion in Eq. (8.41), the spectral SLE representation of significantly differing posteriors requires higher order corrections. Otherwise put, SLEs are expected to perform better for posteriors that only slightly update or perturb the prior.

8.4.8 Change of the reference density

As just discussed, a major drawback of SLEs is their dependency on the prior π as the reference density function. The errors are minimized and measured with respect to the prior and the posterior is represented as correction of the standard reference. In case that high-order corrections are required, SLEs also suffer from the curse of dimensionality. While fully maintaining the advantages of the spectral problem formulation, these shortcomings can be remedied through the introduction of an *auxiliary density* g over the prior support $\mathcal{D}_{\boldsymbol{x}}$ that would optimally mimic the posterior in some sense. This *reference density change* allows for the construction of auxiliary expansions that are more accurate with respect to the posterior and for a more convenient series expansions of the joint posterior density. It is analogous to the adjustment of the integration weight in importance sampling, where the average over a distribution is replaced by a weighted average over another ancillary distribution. An iterative use of this reference change naturally allows for adaptive SLE approaches.

Given that $g(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$, one may define the auxiliary quantity $\mathcal{G} = \mathcal{L}\pi/g$. Under the additional assumption that $\mathcal{G} \in L^2_g(\mathcal{D}_{\boldsymbol{x}})$, one can expand this quantity in terms of polynomials $\{\Psi^g_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^M}$ that are orthogonal with respect to the auxiliary reference. Analogous to the expansion of $\mathcal{L} \in L^2_\pi(\mathcal{D}_{\boldsymbol{x}})$ in Eqs. (8.36) and (8.37), this is

$$\mathcal{G} = \frac{\mathcal{L}\pi}{g} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} b^g_{\boldsymbol{\alpha}} \Psi^g_{\boldsymbol{\alpha}}, \quad \text{with} \tag{8.51}$$

$$b^g_{\boldsymbol{\alpha}} = \langle \mathcal{G}, \Psi^g_{\boldsymbol{\alpha}} \rangle_{L^2_g} = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{G}(\boldsymbol{x}) \Psi^g_{\boldsymbol{\alpha}}(\boldsymbol{x})\, g(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{L}(\boldsymbol{x}) \Psi^g_{\boldsymbol{\alpha}}(\boldsymbol{x})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{8.52}$$

The coefficients of this *auxiliary SLE* (aSLE) are denoted as $\{b^g_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^M}$. They equal the projections $b^g_{\boldsymbol{\alpha}} = \langle \mathcal{L}, \Psi^g_{\boldsymbol{\alpha}}\rangle_{L^2_\pi}$ of the likelihood onto the polynomials $\Psi^g_{\boldsymbol{\alpha}}$. Note that if the new reference $g = \pi$ equals the prior, then $\mathcal{G} = \mathcal{L}$ is simply the likelihood and the formulation remains unchanged. If the density $g = \pi(\cdot|\boldsymbol{y})$ equals the posterior, then the quantity $\mathcal{G} = Z$ equals the model evidence. In this case the aSLE $\mathcal{G} = b^g_{\mathbf{0}}\Psi^g_{\mathbf{0}} = b^g_{\mathbf{0}}$ is a constant with a single nonzero term. If $g \approx \pi(\cdot|\boldsymbol{y})$ only applies in an approximate sense, then one may still speculate that the aSLE is sparser than the corresponding SLE.

As in importance sampling, one can then rewrite the expectation values under π in Eqs. (8.2) and (8.7) as expectations under g. Similar to Eq. (8.40), the model evidence then emerges again as the zeroth expansion term

$$Z = \int_{\mathcal{D}_{\boldsymbol{x}}} \mathcal{G}(\boldsymbol{x})\, g(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = b^g_{\mathbf{0}}. \tag{8.53}$$

Let $h = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} c^g_{\boldsymbol{\alpha}}\Psi^g_{\boldsymbol{\alpha}}$ with $c^g_{\boldsymbol{\alpha}} = \langle h, \Psi^g_{\boldsymbol{\alpha}}\rangle_{L^2_g}$ be the auxiliary expansion of a QoI. Similar to Eq. (8.43), for general QoI posterior expectations one may then write

$$\mathbb{E}[h(\boldsymbol{X})|\boldsymbol{y}] = \frac{1}{Z}\int_{\mathcal{D}_{\boldsymbol{x}}} h(\boldsymbol{x})\mathcal{G}(\boldsymbol{x})\, g(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \frac{1}{b^g_{\mathbf{0}}}\sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} c^g_{\boldsymbol{\alpha}} b^g_{\boldsymbol{\alpha}}. \tag{8.54}$$

In accordance with the aSLE in Eqs. (8.51) and (8.52) the joint density of the posterior distribution is obtained as the asymptotic series

$$\pi(\boldsymbol{x}|\boldsymbol{y}) = \frac{\mathcal{G}(\boldsymbol{x})g(\boldsymbol{x})}{Z} = \frac{1}{b^g_{\mathbf{0}}}\left(\sum_{\boldsymbol{\alpha}\in\mathbb{N}^M} b^g_{\boldsymbol{\alpha}}\Psi^g_{\boldsymbol{\alpha}}(\boldsymbol{x})\right) g(\boldsymbol{x}) = g(\boldsymbol{x})\left(1 + \sum_{\boldsymbol{\alpha}\in\mathbb{N}^M\setminus\{\mathbf{0}\}} \frac{b^g_{\boldsymbol{\alpha}}}{b^g_{\mathbf{0}}}\Psi^g_{\boldsymbol{\alpha}}(\boldsymbol{x})\right). \tag{8.55}$$

As opposed to Eq. (8.41) where the posterior density is represented around the prior π, in Eq. (8.55) the posterior is expanded about the new reference g. If the latter resembles the posterior adequately well, the formulation only calls for small corrections.

8.5 Numerical examples

Next, the potential and the difficulties of the theory presented in the preceding section are investigated. The goal is to give a proof of concept for the basic feasibility of spectral Bayesian inference. It is verified that the theory can be successfully applied in practice and further insight into its functioning is obtained. Moreover, it is learned about its current shortcomings. Four instructive calibration problems from classical statistics and inverse modeling are solved for these purposes. The analysis is confined to problems with low-dimensional parameter spaces. First, the mean value of a normal distribution is inferred with random data under a conjugate normal prior. Second, the mean and standard deviation of a normal distribution are fitted for a joint prior with independent and uniform marginals. Third, an inverse heat conduction problem in two spatial dimensions with two unknowns is solved. Finally, a similar thermal problem with six unknowns is considered. Synthetically created pseudo-data are used in all these example applications.

As it turns out, one can gain valuable insights into the characteristics of likelihood expansions and posterior emulators by way of comparison. Therefore, the analyses for the first three examples proceed analogously. For rich experimental designs, the convergence behavior of high-degree SLEs is studied by reference to the LOO error. More importantly, the capability of lower-degree SLEs to accurately capture the posterior QoI-expectations is explored for scarcer experimental designs. Eventually, aSLE-based posterior surrogates are investigated in order to mitigate the curse of dimensionality. All results are compared to reference solutions. Where possible, the exact solutions from a conjugate Bayesian analysis are used to this effect. Otherwise, corresponding approximations are computed via classical MCMC sampling.

The uncertainty quantification platform UQLab [114, 115] is used throughout the numerical demonstrations. It provides a flexible environment for the uncertainty analysis of engineering systems, e.g. for uncertainty propagation. In this context it ships with a range of regression tools that allow one to easily compute PCEs. These tools can be directly applied to the likelihood function in order to compute SLEs. OLS is employed as the standard solving routine in the following examples.

8.5.1 1D normal fitting

First of all, we consider the problem of fitting a Gaussian distribution $\mathcal{N}(y_i|\mu,\sigma^2)$ to random realizations y_i with $i = 1,\dots,N$. The goal is to estimate the unknown mean μ whereas the standard deviation σ is assumed to

be already known. Given a Gaussian prior, this one-dimensional normal model with known variance exhibits a Gaussian posterior density. Moreover, a closed-form expression for the model evidence can be derived. Since this offers the possibility of comparing the SLE results with analytical solutions, this simple statistical model is used as a first SLE testbed. Let the data $\boldsymbol{y} = (y_1, \dots, y_N)^\top$ be comprised of N independent samples from the normal distribution. For the observational model one may then write

$$\boldsymbol{Y}|\mu \sim \prod_{i=1}^{N} \mathcal{N}(y_i|\mu, \sigma^2), \quad \text{with known } \sigma^2. \tag{8.56}$$

Consequently, the likelihood function can be simply written as $\mathcal{L}(\mu) = \prod_{i=1}^{N} \mathcal{N}(y_i|\mu, \sigma^2)$. A Bayesian prior distribution $\pi(\mu)$ captures the epistemic uncertainty of the true value of μ before the data analysis. For the posterior distribution, that aggregates the information about the unknown after the data have been analyzed, one then has $\pi(\mu|\boldsymbol{y}) = Z^{-1}\mathcal{L}(\mu)\pi(\mu)$.

The conjugate prior for the data model in Eq. (8.56) is a Gaussian $\pi(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$. Its mean $\mu_0 = \mathbb{E}[\mu]$ and variance $\sigma_0^2 = \mathrm{Var}[\mu]$ have to be conveniently specified by the experimenter and data analyst. This prior choice ensures that the posterior is a Gaussian $\pi(\mu|\boldsymbol{y}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$ whose parameters $\mu_N = \mathbb{E}[\mu|\boldsymbol{y}]$ and $\sigma_N^2 = \mathrm{Var}[\mu|\boldsymbol{y}]$ are easily found as

$$\mu_N = \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1} \left(\frac{\mu_0}{\sigma_0^2} + \frac{N\overline{y}}{\sigma^2}\right), \quad \sigma_N^2 = \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1}. \tag{8.57}$$

Here, $\overline{y} = N^{-1}\sum_{i=1}^{N} y_i$ is the empirical sample mean of the data. Likewise, an explicit expression for the model evidence $Z = \int_{\mathbb{R}} (\prod_{i=1}^{N} \mathcal{N}(y_i|\mu, \sigma^2))\, \mathcal{N}(\mu|\mu_0, \sigma_0^2)\, \mathrm{d}\mu$ can be derived. Let $\overline{y^2} = N^{-1}\sum_{i=1}^{N} y_i^2$ denote the sample mean of the squared observations. A straightforward calculation based on simple algebra and a Gaussian integral then yields

$$Z = \sigma_0^{-1} \left(\sigma\sqrt{2\pi}\right)^{-N} \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{\mu_0^2}{\sigma_0^2} + \frac{N\overline{y^2}}{\sigma^2} - \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1} \left(\frac{\mu_0}{\sigma_0^2} + \frac{N\overline{y}}{\sigma^2}\right)^2\right)\right). \tag{8.58}$$

For the following computer experiment, the parameters of the data distribution in Eq. (8.56) are specified as $\mu = 10$ and $\sigma = 5$, respectively. In the course of the procedure only the mean is treated as an unknown, whereas the standard deviation is assumed to be known. We consider a situation where $N = 10$ samples are randomly drawn from the data distribution. For the numerical experiment, the pseudo-random numbers $\boldsymbol{y} = (8.78, 4.05, 12.58, 3.60, 11.05, 8.70, 20.80, 1.23, 19.36, 12.07)^\top$ are used as synthetic data. The prior distribution is set to be a Gaussian $\pi(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ with $\mu_0 = 11.5$ and $\sigma_0 = 1.5$.

8.5.1.1 Posterior density

In order to better understand the principles of spectral Bayesian inference we now proceed as follows. Spectral expansions $\hat{\mathcal{L}}_p$ of the likelihood function $\mathcal{L}$ defined above are computed and compared for experimental designs of varying size K and polynomial terms of varying degree p. Hermite polynomials are used in combination with an appropriate linear transformation to standardized variables $\xi_\mu \in \mathbb{R}$ with a Gaussian weight function $\mathcal{N}(\xi_\mu|0, 1)$. Accordingly, the unknown can be represented as $\mu = \mu_0 + \sigma_0\xi_\mu$. The experimental designs are one-dimensional Sobol sequences that are appropriately transformed.

First the convergence behavior and the accuracy of the likelihood approximation are analyzed. For a rich experimental design with $K = 5 \times 10^4$, SLEs are computed for an increasing order up to $p = 20$. The normalized empirical error ϵ_{Emp} and the normalized LOO error ϵ_{LOO} are monitored over these computations. While the former can be directly computed according to its definition, the computation of the latter relies on the reformulation in Eq. (8.34). This serves the purpose of assessing the prediction accuracy of the computed SLE as a function of the degree p. The results are plotted in Fig. 8.1. It can be seen how the error estimates approach zero, i.e. the SLE converges to the likelihood function. For $p = 20$ the empirical error amounts to $\epsilon_{\mathrm{Emp}} = 1.05 \times 10^{-12}$ and the LOO amounts to $\epsilon_{\mathrm{LOO}} = 1.82 \times 10^{-10}$. These small error magnitudes show that the likelihood function $\mathcal{L}$ can be indeed spectrally expanded in a Hermite basis.

The functional likelihood approximation $\hat{\mathcal{L}}_p$ provided by the most accurate SLE with $p = 20$ is visualized in Fig. 8.2. Moreover, the plot shows a low-order SLE with $p = 5$ and $K = 1 \times 10^2$ for which the error estimates $\epsilon_{\mathrm{Emp}} = 2.61 \times 10^{-4}$ and $\epsilon_{\mathrm{LOO}} = 8.41 \times 10^{-4}$ are obtained. For the sake of comparison the exact likelihood function $\mathcal{L}$ is shown as well. It can be seen that the SLEs are able to accurately represent the likelihood around its peak, i.e. roughly speaking in the interval $\mu \in [8, 15]$ for $p = 5$ and in $\mu \in [5, 18]$ for $p = 20$. Note that these

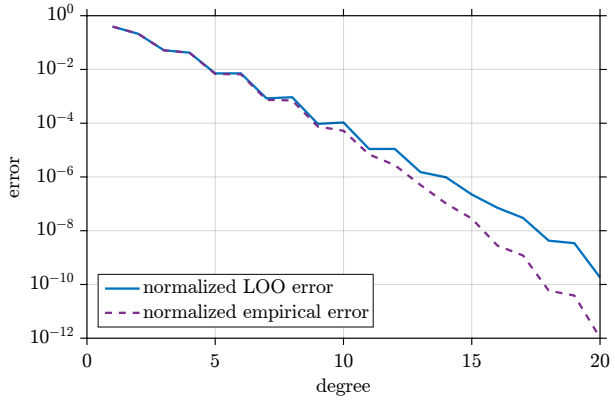


Figure 8.1: 1D normal fitting: Convergence of the SLE.

regions accumulate the largest proportions of the total prior probability mass. Outside of these ranges, however, the SLEs $\hat{\mathcal{L}}_p$ start strongly deviating from $\mathcal{L}$ and taking on negative values. These phenomena can be attributed to an imperfect polynomial cancellation of the finite series approximation of the likelihood in the regions of the parameter space that are only sparsely covered by the experimental design. Indeed, for unbounded parameter spaces it is clearly hopeless to achieve a global net cancellation of a finite polynomial expansion that is necessary in order to emulate the vanishing behavior of the likelihood far from its peaks. The extent to which this impacts on the approximation of the posterior density and its first moments is investigated next.

Expanding the likelihood function is only a means to the end of surrogating the posterior density. Approximations of the posterior density $\pi(\mu|\boldsymbol{y}) \approx b_0^{-1}\hat{\mathcal{L}}_p(\mu)\pi(\mu)$ are computed from the SLEs with $p = 5$ and $p = 20$ through Eqs. (8.39) and (8.40). The results are plotted in Fig. 8.3. In addition to the SLE approximations, the prior density $\pi(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ and the exact solution $\pi(\mu|\boldsymbol{y}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$ from a conjugate analysis based on Eq. (8.57) are shown. The posterior surrogate for $p = 5$ shows minor deviations from the the analytical result, while the approximation for $p = 20$ perfectly matches the true density. It is noted that the discrepancies between $\hat{\mathcal{L}}_p$ and $\mathcal{L}$ shown in Fig. 8.2 are attenuated. The underlying reason is that for large enough $|\mu| \to \infty$ the exponential decay of the Gaussian prior $\pi(\mu) \propto \exp(-(\mu - \mu_0)^2)$ dominates the polynomial increase of $\hat{\mathcal{L}}_p(\mu) = \sum_{\alpha=0}^{p} b_\alpha \Psi_\alpha(\mu)$ in the sense that $\hat{\mathcal{L}}_p(\mu)\pi(\mu) \to 0$. This absorbs the effects of the SLE approximation that is increasingly inadequate for large values of $|\mu|$. In this sense, the prior reference density guards the posterior surrogate against the inadequacies of the SLE. Therefore, the posterior emulation may very well be more accurate than the SLE approximation of the likelihood.

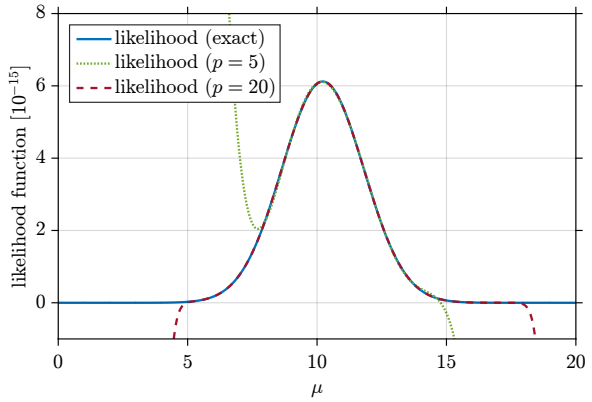


Figure 8.2: 1D normal fitting: Likelihood function.

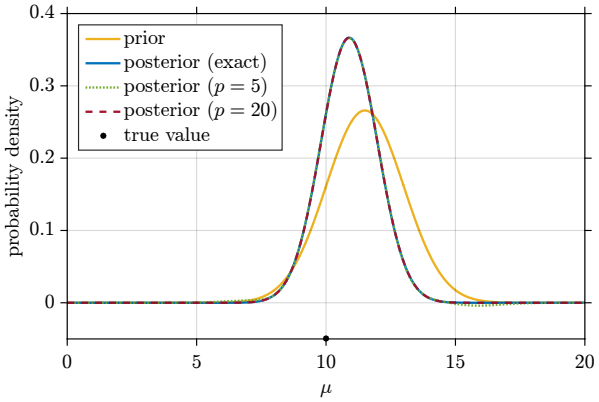


Figure 8.3: 1D normal fitting: Posterior density.

8.5.1.2 Quantities of interest

Commonly one employs posterior means as parameter estimates and posterior standard deviations as measures of the estimation uncertainty. In order to investigate how well one can approximate the model evidence together with these meaningful quantities in spectral Bayesian inference, SLEs are computed for experimental designs of varying size K and for a selection of expansion orders p. The corresponding SLE-based approximations of the model evidence Z, the posterior mean μ_N and the standard deviation σ_N are then computed from Eqs. (8.40),

(8.48) and (8.49). Note that the effects of the transformation to standard variables have to be appropriately taken care of at this place. This happens via Eq. (8.19). The SLE approximations can then be compared to the analytical solutions that are obtained from the conjugate analysis in Eqs. (8.57) and (8.58). In Table 8.2 the results of this procedure are summarized. Note that all the SLE estimates attain admissible values, e.g. the model evidence is non-negative. Furthermore, it is noticed that Z, μ_N and σ_N can be recovered with high accuracy even for very scarce experimental designs and low-order SLEs, say for $K = 1 \times 10^3$ and $p = 10$. It is concluded that, in some sense, the accurate estimation of the model evidence and the first posterior moments require significantly less computational effort than the accurate estimation of the posterior density.

Table 8.2: 1D normal fitting: Statistical quantities.

	K	p	ϵ_{LOO}	Z $[10^{-15}]$	μ_N	σ_N
SLE	1×10^2	5	8.41×10^{-4}	3.71	10.85	0.92
	5×10^2	8	2.49×10^{-4}	3.75	10.91	1.14
	1×10^3	10	2.58×10^{-5}	3.74	10.90	1.07
	5×10^3	12	8.21×10^{-6}	3.74	10.89	1.09
	1×10^4	15	3.84×10^{-7}	3.74	10.89	1.09
	5×10^4	20	1.82×10^{-10}	3.74	10.89	1.09
Exact results				3.74	10.89	1.09

8.5.2 2D normal fitting

Next, we consider the problem of fitting both the unknown mean μ and the standard deviation σ of a Gaussian distribution $\mathcal{N}(y_i|\mu,\sigma^2)$. A number of independent samples y_i with $i = 1,\dots,N$ from the normal distribution constitute the available data $\boldsymbol{y} = (y_1,\dots,y_N)^\top$. The data model for this situation is written as

$$\boldsymbol{Y}|\mu,\sigma \sim \prod_{i=1}^{N} \mathcal{N}(y_i|\mu,\sigma^2). \tag{8.59}$$

For the likelihood function one then has $\mathcal{L}(\mu,\sigma) = \prod_{i=1}^{N} \mathcal{N}(y_i|\mu,\sigma^2)$. Given a Bayesian prior $\pi(\mu,\sigma)$, the posterior distribution is $\pi(\mu,\sigma|\boldsymbol{y}) = Z^{-1}\mathcal{L}(\mu,\sigma)\pi(\mu,\sigma)$. This distribution aggregates the information about the two unknowns after the data have been analyzed.

The true values of the mean and standard deviation are set as $\mu = 30$ and $\sigma = 5$, respectively. These values are treated as unknowns in the further course of the computer experiment. We consider a situation where $N = 10$ samples are randomly drawn from the distribution in Eq. (8.59). The pseudo-random numbers $\boldsymbol{y} = (31.23, 27.50, 24.91, 25.99, 32.88, 36.41, 27.81, 25.19, 37.96, 34.84)^\top$ are used as synthetic data. We consider an independent prior $\pi(\mu,\sigma) = \pi(\mu)\pi(\sigma)$ with uniform marginals $\pi(\mu) = \mathcal{U}(\mu|\underline{\mu},\overline{\mu})$ and $\pi(\sigma) = \mathcal{U}(\sigma|\underline{\sigma},\overline{\sigma})$ over bounded supports $\mathcal{D}_\mu = [\underline{\mu},\overline{\mu}] = [20, 40]$ and $\mathcal{D}_\sigma = [\underline{\sigma},\overline{\sigma}] = [2, 10]$. As opposed to the conjugate example above, this two-dimensional model does not permit a closed-form expression of the posterior density and the model evidence.

8.5.2.1 Posterior density

Now we proceed analogously to the investigation of the normal model with known variance. Expansions $\hat{\mathcal{L}}_p$ of the likelihood $\mathcal{L}$ are computed and contrasted for different experimental designs of size K and different polynomial orders p. An appropriate linear transformation to uniform standardized variables is applied such that the unknowns are represented as $\mu = (\overline{\mu} - \underline{\mu})/2 \cdot \xi_\mu + (\underline{\mu} + \overline{\mu})/2$ and $\sigma = (\overline{\sigma} - \underline{\sigma})/2 \cdot \xi_\sigma + (\underline{\sigma} + \overline{\sigma})/2$, respectively. Here, $\xi_\mu, \xi_\sigma \in [-1, 1]$ are the corresponding standardized variables with a uniform weight function. Accordingly, tensorized Legendre polynomials form the trial basis. Two-dimensional Sobol sequences are utilized as uniformly space-filling experimental designs.

As before, the speed of convergence and the prediction accuracy of the SLE are analyzed first. The normalized empirical error ϵ_{Emp} and the normalized LOO error ϵ_{LOO} are therefore monitored throughout a series of runs that are conducted for an experimental design of the fixed size $K = 1 \times 10^5$ and for an increasing expansion order up to $p = 50$. In Fig. 8.4 a corresponding plot is shown, where the convergence of the SLE $\hat{\mathcal{L}}_p$ to the likelihood function $\mathcal{L}$ is diagnosed. The reason that ϵ_{Emp} and ϵ_{LOO} do not significantly differ is that the large size of the experimental design prevents overfitting. For $p = 50$ the normalized empirical error and the normalized LOO error are found as $\epsilon_{\text{Emp}} = 5.56 \times 10^{-11}$ and $\epsilon_{\text{LOO}} = 6.05 \times 10^{-11}$, respectively. This shows that the likelihood

function $\mathcal{L}$ can be indeed expanded in the Legendre basis. For the uniform prior distribution that is used here, the normalized SLE errors effectively measure the errors of the posterior density.

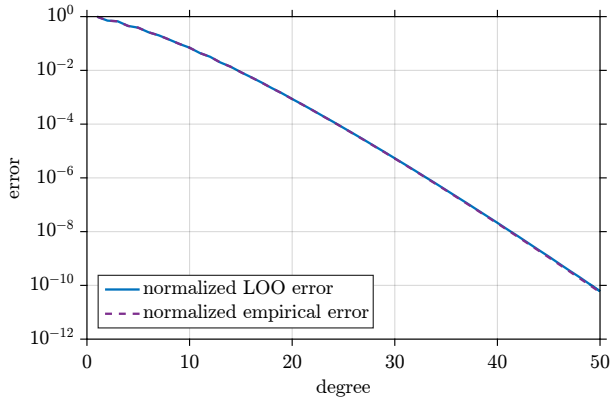
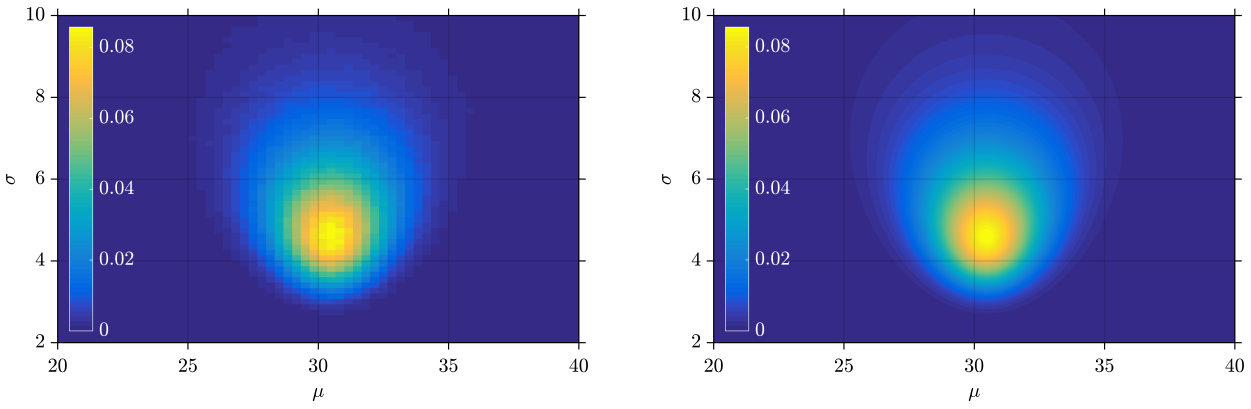


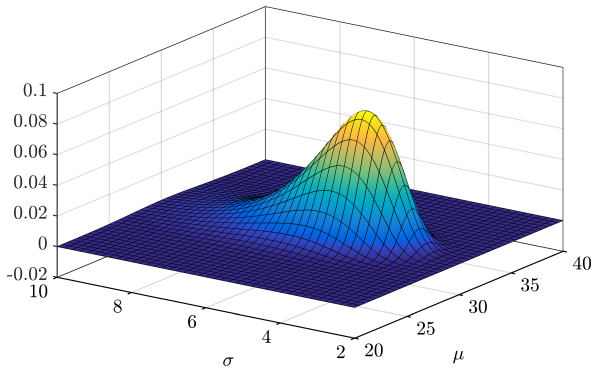
Figure 8.4: 2D normal fitting: Convergence of the SLE.

Now the joint posterior density $\pi(\mu, \sigma|\boldsymbol{y})$ is computed and plotted in Fig. 8.5. For comparison purposes the posterior is sampled by means of MCMC simulation first. A simple random walk Metropolis (RWM) sampler with a Gaussian instrumental distribution is utilized. With this algorithm an unusually large number of 10^7 MCMC samples is drawn from the posterior. This serves the purpose of providing very accurate results that act as references for the SLE-based estimates. In Fig. 8.5(a) a normalized histogram of the obtained RWM sample is shown. Next, the joint posterior density $\pi(\mu, \sigma|\boldsymbol{y}) \approx b_{\mathbf{0}}^{-1} \hat{\mathcal{L}}_p(\mu, \sigma)\pi(\mu, \sigma)$ is computed via Eqs. (8.39) and (8.40). The SLE $\hat{\mathcal{L}}_p(\mu, \sigma)$ with $p = 50$ that features the lowest LOO error is used. In Fig. 8.5(b) the posterior surrogate that arises from the SLE is plotted. For a later comparison with the heat conduction example, in Fig. 8.5(c) the SLE posterior surrogate from Fig. 8.5(b) is plotted again from a different angle. By visual inspection the obvious similarity between the density $\pi(\mu, \sigma|\boldsymbol{y})$ sampled by MCMC and emulated by the SLE is noticed.



(a) MCMC reference sample.

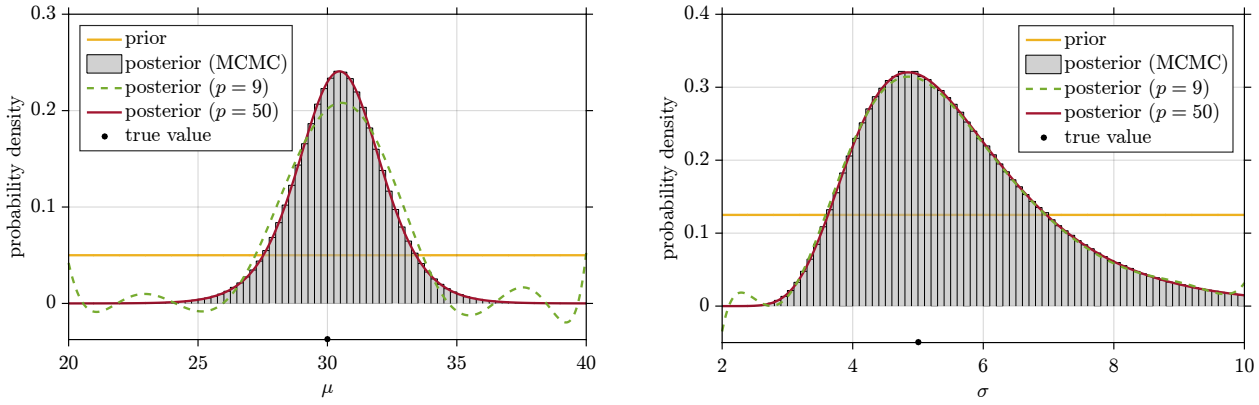
(b) SLE with $p = 50$.



(c) SLE with $p = 50$.

Figure 8.5: 2D normal fitting: Joint posterior.

Now the posterior marginals $\pi(\mu|\boldsymbol{y})$ and $\pi(\sigma|\boldsymbol{y})$ are computed from the joint posterior. On the one hand, samples from the posterior marginals are obtained by restricting the analysis to the corresponding components of the joint MCMC sample. On the other hand, functional approximations of the posterior marginals are extracted based on sub-expansions $\hat{\mathcal{L}}_{\mu,p}(\mu)$ and $\hat{\mathcal{L}}_{\sigma,p}(\sigma)$ of a joint SLE $\hat{\mathcal{L}}_p(\mu,\sigma)$ as in Eqs. (8.44) and (8.45). For the SLEs with $p = 9$ and $p = 50$ the results are visualized in Fig. 8.6. Histogram-based MCMC sample representations and functional SLE approximations of the marginal densities are shown, too. As it can be seen, the marginal posteriors as obtained by MCMC and the SLE with $p = 50$ exactly match each other. For $p = 9$ the posteriors marginals display some wavelike fluctuations in their tails.



(a) Mean value μ.

(b) Standard deviation σ.

Figure 8.6: 2D normal fitting: Posterior marginals.

8.5.2.2 Quantities of interest

Since the posterior density itself is of little inferential use, the model evidence and the first posterior moments are computed for a selection of SLEs with varying size of the experimental design K and degree p. According to Eqs. (8.40) and (8.48) to (8.50), the SLE estimates of these quantities are obtained from the expansions coefficients. In Table 8.3 a summary of the results is given. Compliant with Eq. (8.40) the SLE estimates of the model evidence Z are obtained as the coefficient of the constant expansion term. According to Eqs. (8.48) and (8.49), the SLE estimates of the posterior mean $\mathbb{E}[\mu|\boldsymbol{y}]$ and the standard deviation $\mathrm{Std}[\mu|\boldsymbol{y}] = \mathrm{Var}[\mu|\boldsymbol{y}]^{1/2}$ of the location parameter μ are computed. Likewise, the corresponding estimates for the spread parameter σ follow through a simple postprocessing of the low-order expansion coefficients. The SLE estimates of the linear coefficient of correlation $\rho[\mu,\sigma|\boldsymbol{y}] = \mathrm{Cov}[\mu,\sigma|\boldsymbol{y}]/\mathrm{Std}[\mu|\boldsymbol{y}]/\mathrm{Std}[\sigma|\boldsymbol{y}]$ are computed based on Eq. (8.50). Additionally, the LOO error ϵ_{LOO} is listed to indicate the SLE prediction accuracy. Note that all those estimates comply with the natural bounds and restrictions of the estimated quantities, e.g. the posterior means comply with the prior bounds.

For the sake of comparison, associated results are listed for the simulated MCMC sample, too. These MCMC results are simply obtained as the corresponding sample approximations. The reference estimate of the model evidence is obtained by crude MC simulation instead, i.e. the arithmetic mean of the likelihood is computed for a number of 10^8 independent draws from the prior. It is interesting to note that the SLEs can reproduce the MCMC results for moderate experimental designs and degrees, say for $K = 5 \times 10^3$ and $p = 21$. Even though a large number of input samples and a large polynomial degree is necessary to reproduce the shape of the joint posterior density, significantly smaller experimental designs and polynomial orders suffice to reproduce the first posterior moments.

8.5.3 2D inverse heat conduction

Finally, an inverse heat conduction problem (IHCP) is considered. The heat equation is a partial differential equation (PDE) that describes the distribution and evolution of heat in a system where conduction is the dominant mode of heat transfer. We consider a stationary heat equation of the form

$$\nabla \cdot (\kappa \nabla \tilde{T}) = 0. \tag{8.60}$$

The temperature is denoted as $\tilde{T}$ and the thermal conductivity is denoted as κ. Commonly one is interested in the solution of the boundary value problem that is posed when Eq. (8.60) is satisfied over a physical domain

Table 8.3: 2D normal fitting: Statistical quantities.

	K	p	ϵ_{LOO}	Z $[10^{-14}]$	$\mathbb{E}[\mu\|\boldsymbol{y}]$	$\mathbb{E}[\sigma\|\boldsymbol{y}]$	$\text{Std}[\mu\|\boldsymbol{y}]$	$\text{Std}[\sigma\|\boldsymbol{y}]$	$\rho[\mu,\sigma\|\boldsymbol{y}]$
SLE	5×10^2	5	4.24×10^{-1}	1.19	30.34	5.57	2.03	1.39	0.18
	1×10^3	9	1.19×10^{-1}	1.20	30.39	5.54	2.01	1.41	0.08
	5×10^3	21	9.64×10^{-4}	1.18	30.48	5.56	1.79	1.38	−0.01
	1×10^4	32	5.86×10^{-6}	1.18	30.47	5.56	1.81	1.38	0.00
	5×10^4	45	1.30×10^{-9}	1.18	30.47	5.56	1.81	1.38	−0.00
	1×10^5	50	6.05×10^{-11}	1.18	30.47	5.56	1.81	1.38	−0.00
(MC)MC				1.18	30.47	5.56	1.81	1.38	−0.00

subject to appropriate boundary conditions. We consider the steady state situation in two spatial dimensions. The Euclidean coordinate vector is denoted as $\boldsymbol{r} = (r_1, r_2)^\top$ in the following.

It is dealt with the identification of thermal conductivities of inclusions in a composite material with close-to-surface measurements of the temperature. The setup of the simplified thermal problem is visualized in Fig. 8.7. The thermal conductivity of the background matrix is denoted as κ_0, while the conductivities of the material inclusions are termed as κ_1 and κ_2, respectively. It is assumed that the material properties are not subject to a further spatial variability. At the "top" of the domain a Dirichlet boundary condition $\tilde{T}_1$ is imposed, while at the "bottom" the Neumann boundary condition $q_2 = -\kappa_0 \, \partial\tilde{T}/\partial r_2$ is imposed. Zero heat flux conditions $\partial\tilde{T}/\partial r_1 = 0$ are imposed at the "left" and "right" hand side.

We consider the IHCP that is posed when the thermal conductivities $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)^\top$ are unknown and their inference is intended. With this in mind, a number of N measurements $\boldsymbol{T} = (T(\boldsymbol{r}_1), \ldots, T(\boldsymbol{r}_N))^\top$ of the temperature field at the measurement locations $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)$ is available. The forward model $\mathcal{M}\colon \boldsymbol{\kappa} \mapsto \tilde{\boldsymbol{T}}$ establishes the connection between the data and the unknowns. It formalizes the operation of solving Eq. (8.60) for $\tilde{\boldsymbol{T}}$ as a function of $\boldsymbol{\kappa}$. Measured temperatures $\boldsymbol{T} = \tilde{\boldsymbol{T}} + \boldsymbol{\varepsilon}$ consist of the corresponding model response $\tilde{\boldsymbol{T}} = \mathcal{M}(\boldsymbol{\kappa})$ and a residual term $\boldsymbol{\varepsilon}$. The latter accounts for measurement uncertainty and forward model inadequacy. We consider residuals that are distributed according to a Gaussian $\mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$. In compliance with Eq. (8.9) the likelihood function is given as $\mathcal{L}(\boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{T}|\mathcal{M}(\boldsymbol{\kappa}), \boldsymbol{\Sigma})$. Provided that a prior distribution $\pi(\boldsymbol{\kappa})$ can be elicited, the posterior is given as $\pi(\boldsymbol{\kappa}|\boldsymbol{T}) = Z^{-1}\mathcal{L}(\boldsymbol{\kappa})\pi(\boldsymbol{\kappa})$.

The thermal conductivity of the background matrix is set to $\kappa_0 = 15\,\text{W/m/K}$, while the thermal conductivities of the inclusions are specified as $\kappa_1 = 32\,\text{W/m/K}$ and $\kappa_1 = 28\,\text{W/m/K}$. The material properties of the inclusions are treated as unknowns subsequently. Moreover, the boundary conditions $\tilde{T}_1 = 200\,\text{K}$ and $q_2 = 2000\,\text{W/m}^2$ are imposed. A finite element (FE) model is used to solve a weak form of the governing PDE. The FE solution for the experimental setup described above is shown in Fig. 8.8. We consider a uniform prior distribution $\pi(\boldsymbol{\kappa}) = \pi(\kappa_1)\pi(\kappa_2)$ with independent marginals $\pi(\kappa_1) = \mathcal{U}(\kappa_1|\underline{\kappa}_1, \overline{\kappa}_1)$ and $\pi(\kappa_2) = \mathcal{U}(\kappa_2|\underline{\kappa}_2, \overline{\kappa}_2)$. The prior bounds are chosen as $\underline{\kappa}_1 = \underline{\kappa}_2 = 20\,\text{W/m/K}$ and $\overline{\kappa}_1 = \overline{\kappa}_2 = 40\,\text{W/m/K}$, respectively. A number of $N = 12$ close-to-surface observations is analyzed. Their measurement locations are indicated by the black dots in Fig. 8.7. Independent Gaussian measurement noise with $\boldsymbol{\Sigma} = \sigma_T^2\boldsymbol{1}$ and $\sigma_T = 0.25\,\text{K}$ is considered. Based on this setup, synthetic data are simulated for conducting the computer experiment. This means that the forward model responses $\tilde{\boldsymbol{T}}$ for the true parameter setup are computed and pseudo-random noise is added in order to obtain $\boldsymbol{T}$.

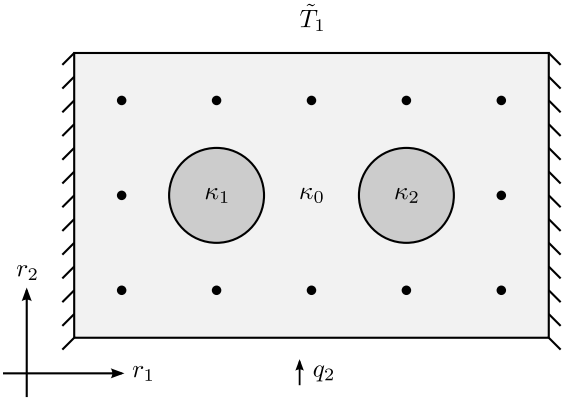


Figure 8.7: 2D IHCP: Heat conduction setup.

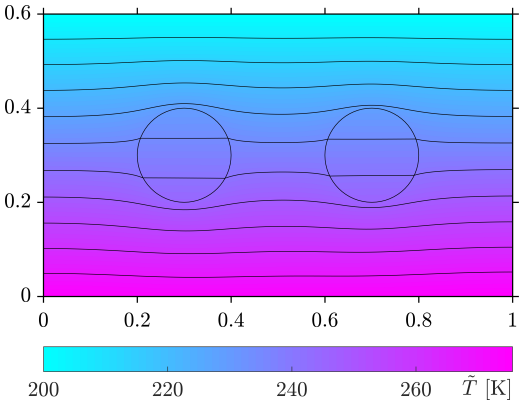


Figure 8.8: 2D IHCP: Steady state solution.

8.5.3.1 Posterior density

The analyses proceed analogously to the preceding section. By comparing the present IHCP and the non-conjugate Gaussian example, that have a two-dimensional parameter space and uniform priors in common, one can gain interesting insight into spectral Bayesian inference. First, the convergence behavior of the SLE is investigated. Spectral expansions $\hat{\mathcal{L}}_p$ of the likelihood $\mathcal{L}$ are therefore computed for an experimental design of size $K = 1 \times 10^5$ and candidate bases with polynomials up to degree $p = 50$. All practical issues are handled analogously to the procedure in the non-conjugate Gaussian example. In Fig. 8.9 the normalized versions of the empirical error ϵ_{Emp} and the LOO error ϵ_{LOO} are shown as a function of p. Comparing these results to Fig. 8.4 reveals that the convergence rate of the SLE $\hat{\mathcal{L}}_p$ is considerably slower than the corresponding one for the Gaussian example. For the SLE with $p = 50$ the error estimates amount to $\epsilon_{\text{Emp}} = 6.26 \times 10^{-4}$ and $\epsilon_{\text{LOO}} = 7.56 \times 10^{-4}$. These errors are around seven orders of magnitude higher than the errors observed for the Gaussian example. The difference in the SLE convergence rate presumably originates from a difference in the underlying likelihood functions and posterior densities. This is now investigated in more detail.

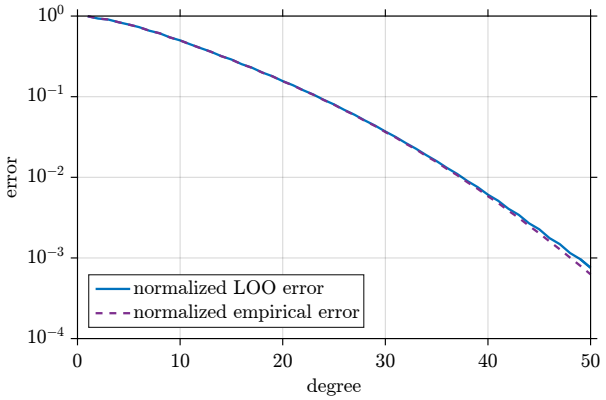
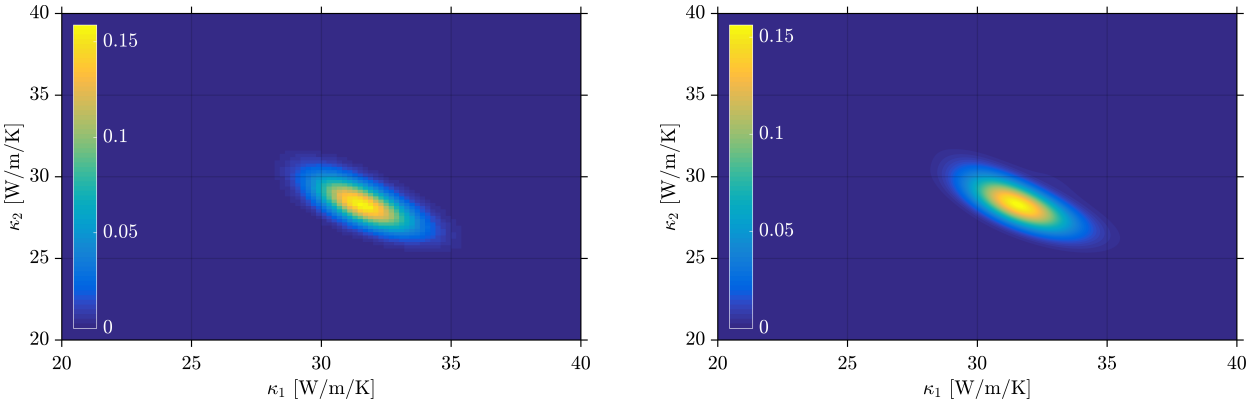


Figure 8.9: 2D IHCP: Convergence of the SLE.

A RWM approximation with 10^7 samples and the SLE-based emulation with $p = 50$ of the posterior density $\pi(\kappa_1, \kappa_2|\boldsymbol{T}) \approx b_{\mathbf{0}}^{-1} \hat{\mathcal{L}}_p(\kappa_1, \kappa_2)\pi(\kappa_1, \kappa_2)$ are depicted in Fig. 8.10. In order to reduce the numerical cost of MCMC sampling, the FE model $\mathcal{M}$ is replaced by a PCE surrogate $\hat{\mathcal{M}}_p$. For $i = 1, \ldots, N$, separate PCEs $\hat{\mathcal{M}}_{i,p}$ of the temperature $\tilde{T}_i = \mathcal{M}_{i,p}(\kappa_1, \kappa_2)$ at the location $\boldsymbol{r}_i$ are fitted as a function of the unknown conductivities. After an appropriate transformation to standardized variables, tensorized Legendre polynomials up to degree $p = 10$ act as the trial basis. Based on an experimental design of the size $K = 10^3$, the LOO errors of the regressions amount to about $\epsilon_{\text{LOO}} \approx 10^{-10}$. Accordingly, the PCE is considered an adequate replacement of the full FE model. Note that it would be also possible to use $\hat{\mathcal{M}}_p$ as a forward model surrogate during the likelihood training runs.

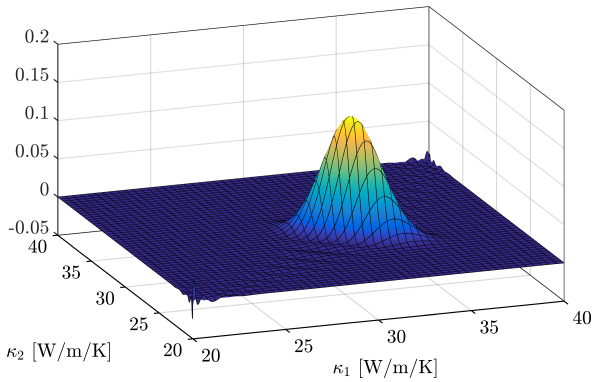
The posteriors in Fig. 8.10 can be compared to the posteriors of the Gaussian example in Fig. 8.5 of the previous section. Relative to the respective prior, the posterior of the thermal problem $\pi(\kappa_1, \kappa_2|\boldsymbol{T})$ contains more information than the posterior of the normal problem $\pi(\mu, \sigma|\boldsymbol{y})$, i.e. the likelihood $\mathcal{L}(\kappa_1, \kappa_2)$ has a slightly more peaked and localized structure than $\mathcal{L}(\mu, \sigma)$. In order to capture these different behaviors nearby and far from the posterior mode, the SLEs $\hat{\mathcal{L}}_p(\kappa_1, \kappa_2)$ and $\hat{\mathcal{L}}_p(\mu, \sigma)$ require a different number of expansions terms. The more localized the posterior modes are with respect to the prior, the more terms are required in order to achieve the cancellation in the tails. Moreover, as opposed to $\pi(\mu, \sigma|\boldsymbol{y})$ the posterior $\pi(\kappa_1, \kappa_2|\boldsymbol{T})$ exhibits a pronounced correlation structure. In turn, this requires non-vanishing interaction terms. As a consequence, the SLE $\hat{\mathcal{L}}_p(\kappa_1, \kappa_2)$ of the IHCP example is less accurate than the SLE $\hat{\mathcal{L}}_p(\mu, \sigma)$ of the Gaussian example. This is also reflected in the fact that the posterior surrogate fluctuates and takes on negative values around the points $[\underline{\kappa}_1, \underline{\kappa}_2]$ and $[\overline{\kappa}_1, \overline{\kappa}_2]$. In order to see this more clearly, the SLE posterior surrogate from Fig. 8.10(b) is plotted again from a different angle in Fig. 8.10(c). A small wavelike posterior structure spans the parameter space between these corners. These artifacts stem from an imperfect polynomial cancellation of the finite series approximation. This stands in contrast to the posterior of the Gaussian example in Fig. 8.5(c) where these phenomena were not observed.

Via Eq. (8.45) the posterior marginals $\pi(\kappa_1|\boldsymbol{T})$ and $\pi(\kappa_2|\boldsymbol{T})$ can be extracted from the joint SLEs. The resulting densities are shown in Fig. 8.11 together with a histogram-based MCMC sample representation. As it can be seen, for $p = 50$ the marginals are captured fairly well, while the moderate-order surrogate for $p = 21$ still exhibits discrepancies at the bounds of the parameter space. The approximation of the posterior marginals by sub-SLEs seems to be more accurate, at least in the sense of the maximum deviation, than the approximation of



(a) MCMC reference sample.

(b) SLE with $p = 50$.



(c) SLE with $p = 50$.

Figure 8.10: 2D IHCP: Joint posterior.

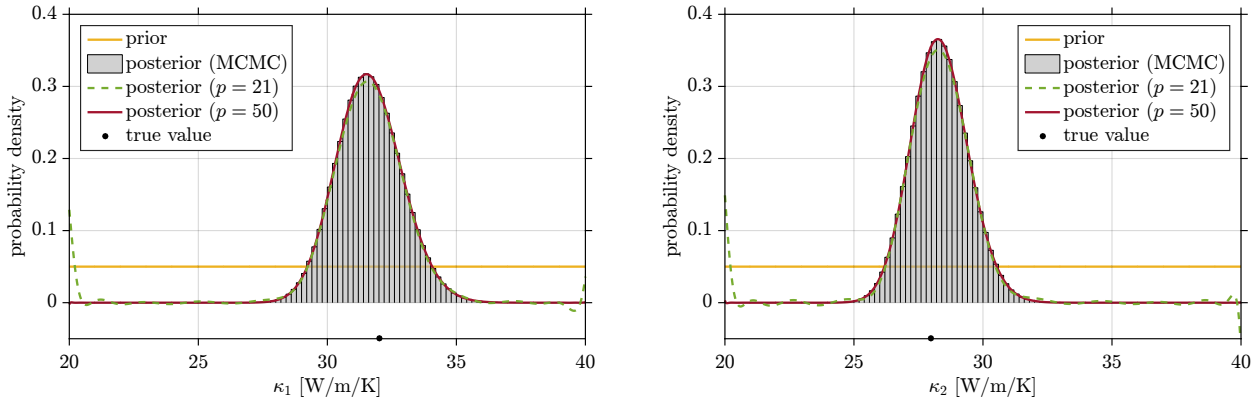
the joint posterior $\pi(\kappa_1, \kappa_2|\boldsymbol{T})$ by the full SLE in Fig. 8.10(c). This phenomenon can be explained through the absence of all non-constant polynomial terms in the variables that are marginalized out.

8.5.3.2 Quantities of interest

Now we investigate how well one can extract the statistically interesting quantities. Results from SLEs with varying K and p are compared with the results from MCMC sampling. A summary of the findings is provided in Table 8.4. The LOO error ϵ_{LOO} of various SLEs is shown together with some basic posterior characteristics obtained by a postprocessing of the SLE coefficients. For $j = 1, 2$ the posterior mean $\mathbb{E}[\kappa_j|\boldsymbol{T}]$ and the standard deviation $\text{Std}[\kappa_j|\boldsymbol{T}] = \text{Var}[\kappa_j|\boldsymbol{T}]^{1/2}$ of the posterior distribution are given in physical units of W/m/K. In addition, the model evidence Z and the linear coefficient of correlation $\rho[\kappa_1, \kappa_2|\boldsymbol{T}] = \text{Cov}[\kappa_1, \kappa_2|\boldsymbol{T}]/\text{Std}[\kappa_1|\boldsymbol{T}]/\text{Std}[\kappa_2|\boldsymbol{T}]$ are specified. In comparison to Table 8.3, where the results for the non-conjugate normal example are listed, the SLE results for the IHCP match their MCMC counterparts less accurately. Nevertheless, it can be observed that the lowest-degree quantities of inferential interest can be extracted with a comparably small experimental design and relatively low number of regressors, say with $K = 1 \times 10^4$ and $p = 29$. Note that all the estimates attain admissible values.

Table 8.4: 2D IHCP: Statistical quantities.

	K	p	ϵ_{LOO}	Z $[10^{-1}]$	$\mathbb{E}[\kappa_1\|\boldsymbol{T}]$	$\mathbb{E}[\kappa_2\|\boldsymbol{T}]$	$\text{Std}[\kappa_1\|\boldsymbol{T}]$	$\text{Std}[\kappa_2\|\boldsymbol{T}]$	$\rho[\kappa_1, \kappa_2\|\boldsymbol{T}]$
SLE	5×10^2	5	8.24×10^{-1}	8.45	31.33	28.36	1.74	1.33	0.28
	1×10^3	9	6.08×10^{-1}	7.81	31.40	28.22	2.02	1.53	0.15
	5×10^3	21	1.50×10^{-1}	7.47	31.32	28.13	2.16	1.61	0.34
	1×10^4	29	5.79×10^{-2}	7.21	31.56	28.30	1.61	1.39	−0.05
	5×10^4	35	1.63×10^{-2}	7.18	31.62	28.34	1.24	1.08	−0.75
	1×10^5	50	7.56×10^{-4}	7.18	31.62	28.33	1.26	1.10	−0.68
(MC)MC				7.17	31.62	28.33	1.26	1.09	−0.68



(a) Thermal conductivity κ_1.

(b) Thermal conductivity κ_2.

Figure 8.11: 2D IHCP: Posterior marginals.

8.5.4 6D inverse heat conduction

In the previous sections it was demonstrated that likelihood functions can be indeed spectrally expanded and that the posterior density with its moments can be computed accordingly. For low-dimensional problems the SLE convergence behavior up to a high degree was studied by monitoring the LOO error. It was shown that the expansion error can be arbitrarily reduced by increasing the order of the expansion and adding samples to the experimental design. While this is reassuring to know, it does not help in solving higher-dimensional problems for which the computation of high-order expansions is exacerbated by the curse of dimensionality. Hence, now we want to investigate the applicability of SLEs and aSLEs in an inverse problem of moderate dimension.

An IHCP in two spatial dimensions with six unknown conductivities is considered in this section. The setup of the problem is shown in Fig. 8.12. The $M = 6$ unknown conductivities $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_6)^\top$ are inferred with $N = 20$ noisy measurements $\boldsymbol{T} = (T_1, \dots, T_{20})^\top$ of the temperature field $\tilde{T}$. We set $\kappa_0 = 30\,\mathrm{W/m/K}$ and $\boldsymbol{\kappa} = (20, 24, \dots, 40)^\top\,\mathrm{W/m/K}$. The prior is set to a multivariate lognormal distribution $\pi(\boldsymbol{\kappa}) = \prod_{i=1}^{6} \pi(\kappa_i)$ with independent marginals $\pi(\kappa_i) = \mathcal{LN}(\kappa_i|\mu_0, \sigma_0^2)$ with $\mu_0 = 30\,\mathrm{W/m/K}$ and $\sigma_0 = 6\,\mathrm{W/m/K}$. These parameters describe the mean $\mu_0 = \mathbb{E}[\kappa_i]$ and standard deviation $\sigma_0 = \mathrm{Std}[\kappa_i]$ of the lognormal prior. They are related to the parameters of the associated normal distribution $\mathcal{N}(\log(\kappa_i)|\lambda_0, \varsigma_0^2)$ via $\mu_0 = \exp(\lambda_0 + \varsigma_0^2/2)$ and $\sigma_0^2 = (\exp(\varsigma_0^2) - 1)\exp(2\lambda_0 + \varsigma_0^2)$. Otherwise than that, the problem setup is exactly as described in the previous section, i.e. the likelihood function is given as $\mathcal{L}(\boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{T}|\mathcal{M}(\boldsymbol{\kappa}), \boldsymbol{\Sigma})$. In accordance with this setup, in the following synthetic data are simulated and analyzed in order to compute the joint posterior $\pi(\boldsymbol{\kappa}|\boldsymbol{T}) = Z^{-1}\mathcal{L}(\boldsymbol{\kappa})\pi(\boldsymbol{\kappa})$.

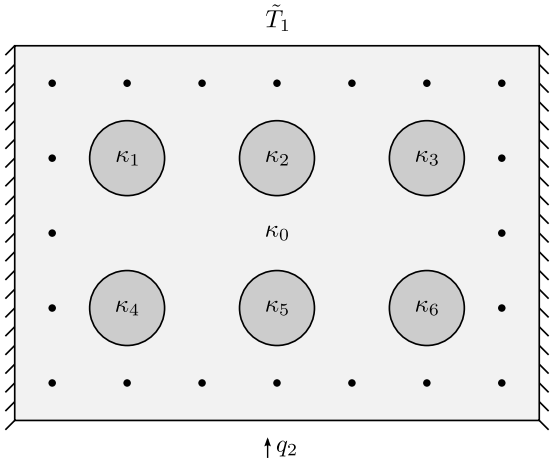


Figure 8.12: 6D IHCP: Heat conduction setup.

8.5.4.1 Posterior density

The unknowns are represented as $\kappa_i = \exp(\lambda_0 + \varsigma_0\xi_i)$ in terms of the standardized variables $\xi_i \in \mathbb{R}$ with Gaussian weight functions $\mathcal{N}(\xi_i|0, 1)$. A spectral expansion $\hat{\mathcal{L}}_p$ in tensorized Hermite polynomials is

then computed for $p = 5$ and $K = 5 \times 10^4$. The errors of the likelihood approximation are estimated as $\epsilon_{\text{Emp}} = 8.81 \times 10^{-1}$ and $\epsilon_{\text{LOO}} = 9.14 \times 10^{-1}$. As compared to the low-dimensional examples that were studied before, these are large errors. An auxiliary reference density $g(\boldsymbol{\kappa}) = \prod_{i=1}^{6} g(\kappa_i)$ is then constructed as a multivariate lognormal with independent marginals $g(\kappa_i) = \mathcal{LN}(\kappa_i|\mu_i, \sigma_i^2)$. The parameters of the latter are chosen as the means $\mu_i = \mathbb{E}[\kappa_i|\boldsymbol{T}]$ and standard deviations $\sigma_i = \text{Std}[\kappa_i|\boldsymbol{T}]$ of the posterior surrogate corresponding to the coefficients of SLE $\hat{\mathcal{L}}_p$. We remark that this is a simple two-step procedure and that a more refined usage of the reference change would certainly lead to more sophisticated approaches. Subsequently, an aSLE $\hat{\mathcal{G}}_p$ with $p = 5$ and $K = 5 \times 10^4$ is computed. The errors amount to $\epsilon_{\text{Emp}} = 4.81 \times 10^{-1}$ and $\epsilon_{\text{LOO}} = 6.24 \times 10^{-1}$. Notwithstanding that these errors are smaller than the corresponding errors of the SLE, they are still large as compared to the previous examples. Since these errors are now measured with respect to the auxiliary density which is expectedly closer to the true posterior than the prior is, the aSLE presumably leads to a more accurate posterior surrogate.

From the previously computed SLE $\hat{\mathcal{L}}(\boldsymbol{\kappa})$ and the aSLE $\hat{\mathcal{G}}_p(\boldsymbol{\kappa})$ approximations of the joint posterior density are computed via Eqs. (8.39) and (8.55). The obtained surrogates $\pi(\boldsymbol{\kappa}|\boldsymbol{T}) \approx \hat{\mathcal{L}}_p(\boldsymbol{\kappa})\pi(\boldsymbol{\kappa})/b_{\mathbf{0}}$ and $\pi(\boldsymbol{\kappa}|\boldsymbol{T}) \approx \hat{\mathcal{G}}_p(\boldsymbol{\kappa})g(\boldsymbol{\kappa})/b_{\mathbf{0}}^g$ are now compared to each other. We start with the one-dimensional marginals that can be compiled by collecting terms from the full expansions based on Eq. (8.45). For $j = 1, \dots, 6$ the marginals $\pi(\kappa_j|\boldsymbol{T})$ that are extracted that way are shown in Fig. 8.13. The marginal priors $\pi(\kappa_j)$ and the auxiliary densities $g(\kappa_j)$ are shown, too. While the marginals that are taken from the SLE slightly deviate from their MCMC counterparts, the marginals based on the aSLE match their references perfectly well. The reason is that the posterior can be easier represented as a small adjustment of the auxiliary density than as a large correction to the prior. Thus, with the same expansion order the posterior is more accurately represented through the aSLE than through the SLE. Regarding the size of the error estimates, it is surprising that the marginals can be retrieved that well with the aSLE. Even though the SLE-based posterior approximations can hardly be interpreted as proper probability densities, i.e. they conspicuously take on negative values, the moments are recovered sufficiently well for the construction of the auxiliary reference density.

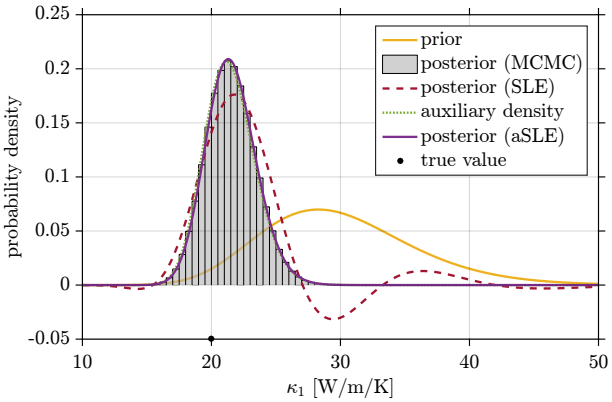
On the basis of Eq. (8.47) the two-dimensional posterior marginals $\pi(\kappa_j, \kappa_k|\boldsymbol{T})$ can be constructed from the full expansions. For $j = 3$ and $k = 4$ the posterior marginal for the SLE $\hat{\mathcal{L}}_p$ is shown in Fig. 8.14(a). The same two-dimensional distribution is depicted in Fig. 8.14(b) for the aSLE $\hat{\mathcal{G}}_p$. A histogram of the MCMC sample is provided in Fig. 8.14(c) as a reference. As already found in Figs. 8.13(c) and 8.13(d) for instance, in Fig. 8.14 the aSLE-based surrogate appears to be almost exact whereas the SLE-based one is flattened out. Since the aSLE captures the true posterior density more accurately than the SLE, we expect similar findings for the posterior moments.

8.5.4.2 Quantities of interest

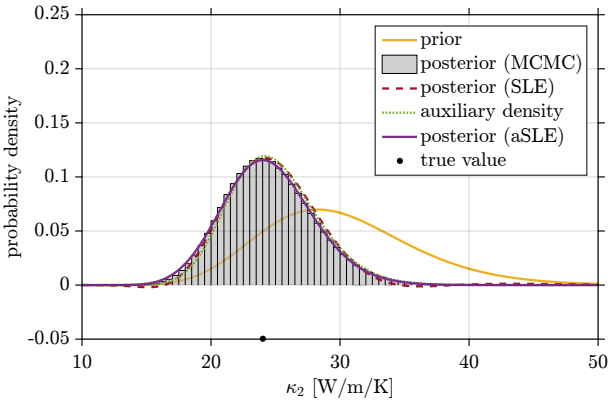
Finally we compute the model evidence and the first posterior moments with the aid of Eqs. (8.40) and (8.53) and Eqs. (8.48) to (8.50). For the aSLE $\hat{\mathcal{G}}_p$ the analysis proceeds analogously to the SLE $\hat{\mathcal{L}}_p$. In Table 8.5 a summary of the results is given. As it can be taken from the table, the aSLE consistently gives more accurate estimates of the reference values. This fulfills our earlier expectations. Regarding the inaccuracy of the SLE-based posterior marginals and the concerns about interpreting them as probability densities, the quality of the SLE-based estimates of the moments surpasses our expectations. In particular, the estimated standard deviations are more accurate than the surrogate marginals suggest, e.g. the ones shown in Figs. 8.13(a) and 8.13(f). Similar as for the posterior density, we have to conclude that the normalized LOO error does not give conclusive information about the accuracy of the first posterior moments. Nevertheless, it is remarked that the use of resampling methods still ensures a robust fit, i.e. it protects against overfitting.

8.6 Concluding remarks

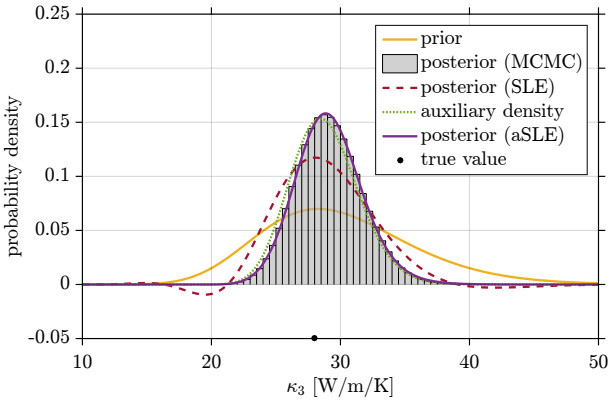
A spectral approach to Bayesian inference that focuses on the surrogate modeling of the posterior density was devised. The likelihood was expanded in terms of polynomials that are orthogonal with respect to the prior weight. Ensuing from this spectral likelihood expansion (SLE), the joint posterior density was expressed as the prior that acts the reference density times a polynomial correction term. The normalization factor of the posterior emerged as the zeroth SLE coefficient and the posterior marginals were shown to be easily accessible through sub-expansions of the SLE. Closed-form expressions for the first posterior moments in terms of the low-order spectral coefficients were given. Posterior uncertainty propagation through general quantities of interest was established via a postprocessing of the higher-order coefficients. The semi-analytic reformulation of Bayesian inference was founded on the theory and practice of metamodeling based on polynomial chaos



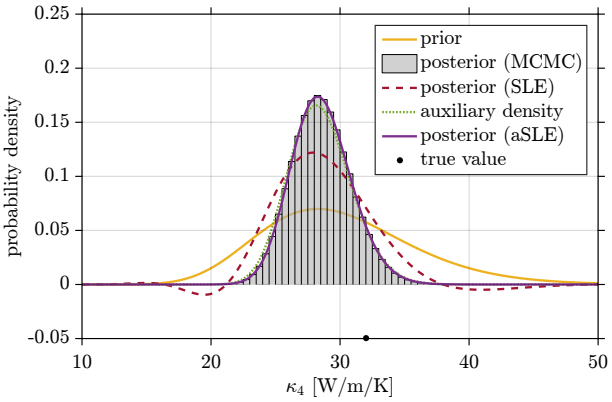
(a) Thermal conductivity κ_1.



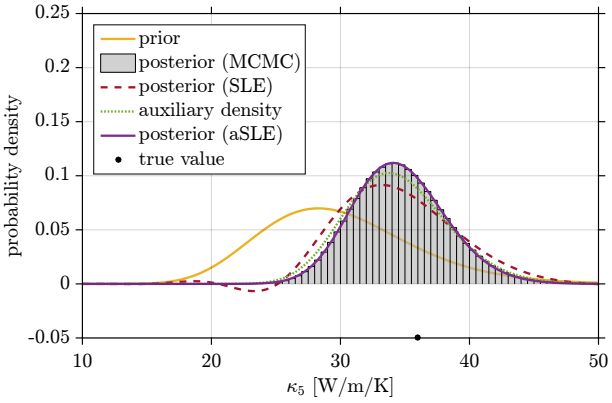
(b) Thermal conductivity κ_2.



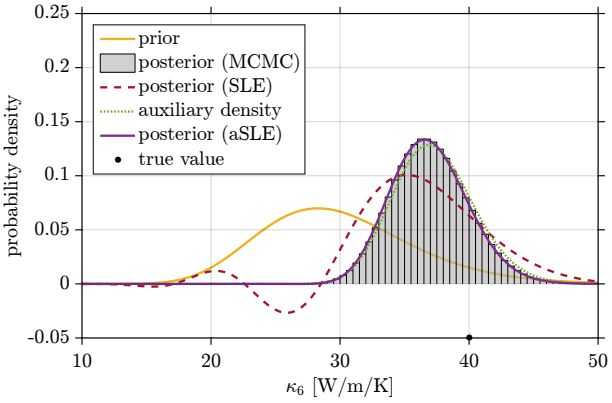
(c) Thermal conductivity κ_3.



(d) Thermal conductivity κ_4.

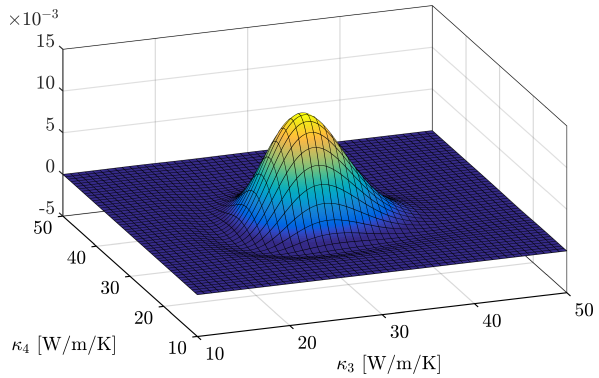


(e) Thermal conductivity κ_5.

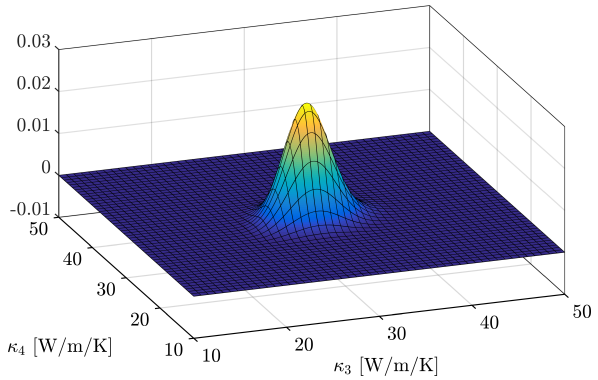


(f) Thermal conductivity κ_6.

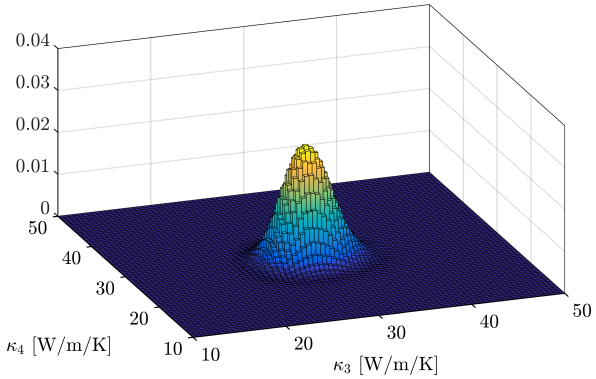
Figure 8.13: 6D IHCP: Posterior marginals.



(a) SLE with $p = 5$.



(b) aSLE with $p = 5$.



(c) MCMC reference sample.

Figure 8.14: 6D IHCP: Posterior marginals.

Table 8.5: 6D IHCP: Statistical quantities.

	Z $[10^{-3}]$	$\mathbb{E}[\kappa_1\|\boldsymbol{T}]$	$\mathbb{E}[\kappa_2\|\boldsymbol{T}]$	$\mathbb{E}[\kappa_3\|\boldsymbol{T}]$	$\mathbb{E}[\kappa_4\|\boldsymbol{T}]$	$\mathbb{E}[\kappa_5\|\boldsymbol{T}]$	$\mathbb{E}[\kappa_6\|\boldsymbol{T}]$
SLE	4.04	21.42	24.86	28.79	28.45	34.43	37.27
aSLE	3.68	21.53	24.48	29.16	28.57	34.59	36.95
(MC)MC	3.65	21.52	24.57	29.11	28.56	34.64	37.00
	$\mathrm{Std}[\kappa_1\|\boldsymbol{T}]$	$\mathrm{Std}[\kappa_2\|\boldsymbol{T}]$	$\mathrm{Std}[\kappa_3\|\boldsymbol{T}]$	$\mathrm{Std}[\kappa_4\|\boldsymbol{T}]$	$\mathrm{Std}[\kappa_5\|\boldsymbol{T}]$	$\mathrm{Std}[\kappa_6\|\boldsymbol{T}]$	$\rho[\kappa_1,\kappa_2\|\boldsymbol{T}]$
SLE	1.95	3.43	2.63	2.43	3.96	3.13	−0.40
aSLE	1.94	3.56	2.61	2.33	3.62	2.99	−0.44
(MC)MC	1.93	3.48	2.56	2.31	3.64	3.00	−0.47
	$\rho[\kappa_1,\kappa_3\|\boldsymbol{T}]$	$\rho[\kappa_1,\kappa_4\|\boldsymbol{T}]$	$\rho[\kappa_1,\kappa_5\|\boldsymbol{T}]$	$\rho[\kappa_1,\kappa_6\|\boldsymbol{T}]$	$\rho[\kappa_2,\kappa_3\|\boldsymbol{T}]$	$\rho[\kappa_2,\kappa_4\|\boldsymbol{T}]$	$\rho[\kappa_2,\kappa_5\|\boldsymbol{T}]$
SLE	0.19	−0.39	−0.28	0.05	−0.40	−0.18	−0.30
aSLE	−0.01	−0.29	−0.03	0.10	−0.48	−0.17	−0.28
(MC)MC	−0.02	−0.32	−0.03	0.09	−0.48	−0.17	−0.31
	$\rho[\kappa_2,\kappa_6\|\boldsymbol{T}]$	$\rho[\kappa_3,\kappa_4\|\boldsymbol{T}]$	$\rho[\kappa_3,\kappa_5\|\boldsymbol{T}]$	$\rho[\kappa_3,\kappa_6\|\boldsymbol{T}]$	$\rho[\kappa_4,\kappa_5\|\boldsymbol{T}]$	$\rho[\kappa_4,\kappa_6\|\boldsymbol{T}]$	$\rho[\kappa_5,\kappa_6\|\boldsymbol{T}]$
SLE	−0.09	−0.00	0.22	−0.22	−0.20	0.24	−0.11
aSLE	−0.13	0.11	−0.02	−0.32	−0.24	0.13	−0.24
(MC)MC	−0.16	0.10	−0.03	−0.34	−0.26	0.12	−0.24

expansions. This allows one to compute the SLE coefficients by solving a linear least squares problem. An analysis of the advantages and disadvantages of the proposed method eventually motivated a change of the reference density. While the expansion of the posterior in terms of the prior may require substantial modifications, its representation with respect to an auxiliary density many only require minor tweaks.

The possibilities and difficulties that arise from the problem formulation were exhaustively discussed and numerically demonstrated. Fitting a parametric distribution to random data and identifying the thermal properties of a composite material served as benchmark problems. These numerical experiments proved that spectral Bayesian inference works in principle and they provided insight into the mechanisms involved. The convergence behavior of the SLE was studied based on the leave-one-out error. It was found that high-degree SLEs are necessary in order to accurately represent the likelihood function and the joint posterior density, whereas lower-order SLEs are sufficient in order to extract the low-level quantities of interest. A change of the reference density allowed for reducing the order of the corrections required in order to represent the posterior with respect to the prior. This helped in alleviating the curse of dimensionality to some extent.

In turn, a number of follow-up questions were given rise to. While the leave-one-out error performs well in quantifying the prediction errors of the SLE, it turned out to be of limited use with regard to the errors of the corresponding posterior surrogate and its marginals. A critical question thus relates to a means to assess the errors of these quantities and to diagnose their convergence. This would assist in choosing experimental designs of a sufficient size. Also, it would be desirable to quantify the estimation errors of individual expansion coefficients. This would support the assessment of the efficiency and scalability of the approach and the fair comparison with Monte Carlo, importance and Markov chain Monte Carlo sampling. Another question is whether a constrained optimization problem can be formulated that naturally respects all prior restrictions. This would remedy the potential problem of illegitimate values of the posterior moments. In order to handle a broader spectrum of statistical problems, SLEs would have to be extended to dependent prior distributions and noisy likelihood functions. For increasing the computational efficiency beyond the change of the reference density, it is conceivable to deploy advanced techniques from metamodeling and machine learning. This includes piecewise polynomial models, expansions in a favorable basis and the use of sparsity-promoting regression techniques. Yet another important issue concerns the practical applicability of the presented framework to problems with higher-dimensional parameter spaces. In future research efforts we will try to address the abovementioned issues and to answer this principal question.

Appendices

8.A Univariate polynomials

The main properties of two classical orthogonal families of polynomials were shortly summarized in Table 8.1, i.e. the domain of definition, the associated weight function and the norm. The first six members of these univariate Hermite polynomials $\{H_\alpha\}_{\alpha\in\mathbb{N}}$ and Legendre polynomials $\{P_\alpha\}_{\alpha\in\mathbb{N}}$ are listed in Table 8.6. Higher order members can be defined via recursive or differential relations. These polynomials can be used for the construction of the multivariate polynomial basis $\{\Psi_\alpha\}_{\alpha\in\mathbb{N}}$ in Eq. (8.17). Note that this orthonormal basis is normalized via $\Psi_\alpha = H_\alpha/\sqrt{\alpha!}$ or $\Psi_\alpha = P_\alpha/\sqrt{1/(2\alpha+1)}$.

Table 8.6: Low-order polynomials.

α	$H_\alpha(x)$, $x \in \mathbb{R}$	$P_\alpha(x)$, $x \in [-1,1]$
0	1	1
1	x	x
2	x^2-1	$(3x^2-1)/2$
3	x^3-3x	$(5x^3-3x)/2$
4	x^4-6x^2+3	$(35x^4-30x^2+3)/8$
5	x^5-10x^3+15x	$(63x^5-70x^3+15x)/8$

8.B Low-order QoIs

The representation of six low-order QoIs in terms of the normalized Hermite and Legendre polynomials is given in Table 8.7 below. Those expansions can be used in order to compute the first posterior moments, e.g. as

shown in Eqs. (8.48) to (8.50). Note that the representations in the orthonormal bases directly follow from a change of basis and the substitutions $H_\alpha = \sqrt{\alpha!}\Psi_\alpha$ and $P_\alpha = \sqrt{1/(2\alpha+1)}\Psi_\alpha$.

Table 8.7: Low-order QoIs.

QoI	Hermite expansion	Legendre expansion
1	Ψ_0	Ψ_0
x	Ψ_1	$\Psi_1/\sqrt{3}$
x^2	$\sqrt{2}\Psi_2 + \Psi_0$	$(2\Psi_2/\sqrt{5} + \Psi_0)/3$
x^3	$\sqrt{6}\Psi_3 + 3\Psi_1$	$(2\Psi_3/\sqrt{7} + 3\Psi_1/\sqrt{3})/5$
x^4	$2\sqrt{6}\Psi_4 + 6\sqrt{2}\Psi_2 + 3\Psi_0$	$(8\Psi_4/3 + 20\Psi_2/\sqrt{5} + 7\Psi_0)/35$
x^5	$2\sqrt{30}\Psi_5 + 10\sqrt{6}\Psi_3 + 15\Psi_1$	$(8\Psi_5/\sqrt{11} + 28\Psi_3/\sqrt{7} + 27\Psi_1/\sqrt{3})/63$

References

[1] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2005. DOI: 10.1137/1.9780898717921.

[2] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Applied Mathematical Sciences 160. New York: Springer, 2005. DOI: 10.1007/b138659.

[3] R. C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Computational Science and Engineering. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2014.

[4] T. J. Sullivan. *Introduction to Uncertainty Quantification*. Texts in Applied Mathematics 63. Cham, Switzerland: Springer International Publishing, 2015. DOI: 10.1007/978-3-319-23395-6.

[5] R. Hadidi and N. Gucunski. "Probabilistic Approach to the Solution of Inverse Problems in Civil Engineering". In: *Journal of Computing in Civil Engineering* 22.6 (2008), pp. 338–347. DOI: 10.1061/(ASCE)0887-3801(2008)22:6(338).

[6] J. L. Beck. "Bayesian system identification based on probability logic". In: *Structural Control and Health Monitoring* 17.7 (2010), pp. 825–847. DOI: 10.1002/stc.424.

[7] K.-V. Yuen and S.-C. Kuok. "Bayesian Methods for Updating Dynamic Models". In: *Applied Mechanics Reviews* 64.1, 010802 (2011), pp. 1–18. DOI: 10.1115/1.4004479.

[8] M. Evans and T. Swartz. "Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems". In: *Statistical Science* 10.3 (1995), pp. 254–272. DOI: 10.1214/ss/1177009938.

[9] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1996.

[10] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, eds. *Handbook of Markov Chain Monte Carlo*. Handbooks of Modern Statistical Methods. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2011. DOI: 10.1201/b10905.

[11] J. Beck and S. Au. "Bayesian Updating of Structural Models and Reliability using Markov Chain Monte Carlo Simulation". In: *Journal of Engineering Mechanics* 128.4 (2002), pp. 380–391. DOI: 10.1061/(ASCE)0733-9399(2002)128:4(380).

[12] J. Ching and Y. Chen. "Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection, and Model Averaging". In: *Journal of Engineering Mechanics* 133.7 (2007), pp. 816–832. DOI: 10.1061/(ASCE)0733-9399(2007)133:7(816).

[13] S. H. Cheung and J. L. Beck. "Bayesian Model Updating Using Hybrid Monte Carlo Simulation with Application to Structural Dynamic Models with Many Uncertain Parameters". In: *Journal of Engineering Mechanics* 135.4 (2009), pp. 243–255. DOI: 10.1061/(ASCE)0733-9399(2009)135:4(243).

[14] I. Boulkaibet, L. Mthembu, T. Marwala, M. I. Friswell, and S. Adhikari. "Finite element model updating using the shadow hybrid Monte Carlo technique". In: *Mechanical Systems and Signal Processing* 52–53 (2015), pp. 115–132. DOI: 10.1016/j.ymssp.2014.06.005.

[15] J. B. Nagel and B. Sudret. "Hamiltonian Monte Carlo and Borrowing Strength in Hierarchical Inverse Problems". In: *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 2.3, B4015008 (2016), pp. 1–12. DOI: 10.1061/AJRUA6.0000847.

[16] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. "Combining Field Data and Computer Simulations for Calibration and Prediction". In: *SIAM Journal on Scientific Computing* 26.2 (2004), pp. 448–466. DOI: 10.1137/S1064827503426693.

[17] D. Higdon, J. D. McDonnell, N. Schunck, J. Sarich, and S. M. Wild. "A Bayesian approach for parameter estimation and prediction using a computationally intensive model". In: *Journal of Physics G: Nuclear and Particle Physics* 42.3, 034009 (2015), pp. 1–18. DOI: 10.1088/0954-3899/42/3/034009.

[18] Y. M. Marzouk, H. N. Najm, and L. A. Rahn. "Stochastic spectral methods for efficient Bayesian solution of inverse problems". In: *Journal of Computational Physics* 224.2 (2007), pp. 560–586. DOI: 10.1016/j.jcp.2006.10.010.

[19] Y. M. Marzouk and H. N. Najm. "Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems". In: *Journal of Computational Physics* 228.6 (2009), pp. 1862–1902. DOI: 10.1016/j.jcp.2008.11.024.

[20] Y. Marzouk and D. Xiu. "A Stochastic Collocation Approach to Bayesian Inference in Inverse Problems". In: *Communications in Computational Physics* 6.4 (2009), pp. 826–847. DOI: 10.4208/cicp.2009.v6.p826.

[21] J. T. Ormerod and M. P. Wand. "Explaining Variational Approximations". In: *The American Statistician* 64.2 (2010), pp. 140–153. DOI: 10.1198/tast.2010.09058.

[22] C. W. Fox and S. J. Roberts. "A tutorial on variational Bayesian inference". In: *Artificial Intelligence Review* 38.2 (2012), pp. 85–95. DOI: 10.1007/s10462-011-9236-8.

[23] S. Sun. "A review of deterministic approximate inference techniques for Bayesian machine learning". In: *Neural Computing and Applications* 23.7–8 (2013), pp. 2039–2050. DOI: 10.1007/s00521-013-1445-4.

[24] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233. DOI: 10.1023/A:1007665907178.

[25] T. S. Jaakkola and M. I. Jordan. "Bayesian parameter estimation via variational methods". In: *Statistics and Computing* 10.1 (2000), pp. 25–37. DOI: 10.1023/A:1008932416310.

[26] M. A. Chappell, A. R. Groves, B. Whitcher, and M. W. Woolrich. "Variational Bayesian Inference for a Nonlinear Forward Model". In: *IEEE Transactions on Signal Processing* 57.1 (2009), pp. 223–236. DOI: 10.1109/TSP.2008.2005752.

[27] B. Jin and J. Zou. "Hierarchical Bayesian inference for Ill-posed problems via variational method". In: *Journal of Computational Physics* 229.19 (2010), pp. 7317–7343. DOI: 10.1016/j.jcp.2010.06.016.

[28] T. A. El Moselhy and Y. M. Marzouk. "Bayesian inference with optimal maps". In: *Journal of Computational Physics* 231.23 (2012), pp. 7815–7850. DOI: 10.1016/j.jcp.2012.07.022.

[29] C. Schwab and A. M. Stuart. "Sparse deterministic approximation of Bayesian inverse problems". In: *Inverse Problems* 28.4, 045003 (2012), pp. 1–32. DOI: 10.1088/0266-5611/28/4/045003.

[30] C. Schillings and C. Schwab. "Sparse, adaptive Smolyak quadratures for Bayesian inverse problems". In: *Inverse Problems* 29.6, 065011 (2013), pp. 1–28. DOI: 10.1088/0266-5611/29/6/065011.

[31] J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. 2nd ed. Dover Books on Mathematics. Mineola, New York, USA: Dover Publications, 2001.

[32] D. A. Kopriva. *Implementing Spectral Methods for Partial Differential Equations: Algorithms for Scientists and Engineers*. Scientific Computation. Dordrecht, Netherlands: Springer, 2009. DOI: 10.1007/978-90-481-2261-5.

[33] J. Shen, T. Tang, and L.-L. Wang. *Spectral Methods: Algorithms, Analysis and Applications*. Vol. 41. Springer Series in Computational Mathematics. Springer-Verlag Berlin Heidelberg, 2011. DOI: 10.1007/978-3-540-71041-7.

[34] O. Christensen and K. L. Christensen. *Approximation Theory: From Taylor Polynomials to Wavelets*. Applied and Numerical Harmonic Analysis. Boston, Massachusetts, USA: Birkhäuser, 2004.

[35] R. M. Trigub and E. S. Belinsky. *Fourier Analysis and Approximation of Functions*. Dordrecht, Netherlands: Springer, 2004. DOI: 10.1007/978-1-4020-2876-2.

[36] L. N. Trefethen. *Approximation Theory and Approximation Practice*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2013.

[37] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Classics in Applied Mathematics. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 1995. DOI: 10.1137/1.9781611971217.

[38] Å. Björck. *Numerical Methods for Least Squares Problems*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 1996. DOI: 10.1137/1.9781611971484.

[39] V. N. Vapnik. *The Nature of Statistical Learning Theory*. 2nd ed. Statistics for Engineering and Information Science. New York: Springer, 2000. DOI: 10.1007/978-1-4757-3264-1.

[40] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York: Springer, 2009. DOI: 10.1007/978-0-387-84858-7.

[41] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. "Approximate Bayesian computational methods". In: *Statistics and Computing* 22.6 (2012), pp. 1167–1180. DOI: 10.1007/s11222-011-9288-2.

[42] M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. "Approximate Bayesian Computation". In: *PLoS Computational Biology* 9.1, e1002803 (2013), pp. 1–10. DOI: 10.1371/journal.pcbi.1002803.

[43] G. Taraldsen and B. H. Lindqvist. "Improper Priors Are Not Improper". In: *The American Statistician* 64.2 (2010), pp. 154–158. DOI: 10.1198/tast.2010.09116.

[44] K. P. P. K. "Generalized priors in Bayesian inversion problems". In: *Advances in Water Resources* 36 (2012), pp. 3–10. DOI: 10.1016/j.advwatres.2011.05.005.

[45] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. "A weakly informative default prior distribution for logistic and other regression models". In: *The Annals of Applied Statistics* 2.4 (2008), pp. 1360–1383. DOI: 10.1214/08-AOAS191.

[46] J. A. Fúquene, J. D. Cook, and L. R. Pericchi. "A case for robust Bayesian priors with applications to clinical trials". In: *Bayesian Analysis* 4.4 (2009), pp. 817–846. DOI: 10.1214/09-BA431.

[47] J. B. Nagel and B. Sudret. "Bayesian Multilevel Model Calibration for Inverse Problems Under Uncertainty with Perfect Data". In: *Journal of Aerospace Information Systems* 12.1 (2015), pp. 97–113. DOI: 10.2514/1.I010264.

[48] J. B. Nagel and B. Sudret. "A unified framework for multilevel uncertainty quantification in Bayesian inverse problems". In: *Probabilistic Engineering Mechanics* 43 (2016), pp. 68–84. DOI: 10.1016/j.probengmech.2015.09.007.

[49] S. Jackman. *Bayesian Analysis for the Social Sciences*. Wiley Series in Probability and Statistics. Chichester, West Sussex, UK: John Wiley & Sons, Ltd., 2009. DOI: 10.1002/9780470686621.

[50] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. 3rd ed. Texts in Statistical Science. Boca Raton, Florida, USA: CRC Press, 2014.

[51] A. M. Stuart. "Inverse problems: A Bayesian perspective". In: *Acta Numerica* 19 (2010), pp. 451–559. DOI: 10.1017/S0962492910000061.

[52] O. G. Ernst, B. Sprungk, and H.-J. Starkloff. "Bayesian Inverse Problems and Kalman Filters". In: *Extraction of Quantifiable Information from Complex Systems*. Ed. by S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, K. Ritter, R. Schneider, C. Schwab, and H. Yserentant. Vol. 102. Lecture Notes in Computational Science and Engineering. Cham, Switzerland: Springer International Publishing, 2014, pp. 133–159. DOI: 10.1007/978-3-319-08159-5_7.

[53] R. E. Caflisch. "Monte Carlo and quasi-Monte Carlo methods". In: *Acta Numerica* 7 (1998), pp. 1–49. DOI: 10.1017/S0962492900002804.

[54] S. T. Tokdar and R. E. Kass. "Importance sampling: a review". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.1 (2010), pp. 54–60. DOI: 10.1002/wics.56.

[55] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[56] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97.

[57] M. K. Cowles and B. P. Carlin. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review". In: *Journal of the American Statistical Association* 91.434 (1996), pp. 883–904. DOI: 10.1080/01621459.1996.10476956.

[58] S. P. Brooks and G. O. Roberts. "Convergence assessment techniques for Markov chain Monte Carlo". In: *Statistics and Computing* 8.4 (1998), pp. 319–335. DOI: 10.1023/A:1008820505350.

[59] A. Vehtari and J. Ojanen. "A survey of Bayesian predictive methods for model assessment, selection and comparison". In: *Statistics Surveys* 6 (2012), pp. 142–228. DOI: 10.1214/12-SS102.

[60] J. Beck and K. Yuen. "Model Selection Using Response Measurements: Bayesian Probabilistic Approach". In: *Journal of Engineering Mechanics* 130.2 (2004), pp. 192–203. DOI: 10.1061/(ASCE)0733-9399(2004)130:2(192).

[61] K.-V. Yuen. "Recent developments of Bayesian model class selection and applications in civil engineering". In: *Structural Safety* 32.5 (2010), pp. 338–346. DOI: 10.1016/j.strusafe.2010.03.011.

[62] C. Han and B. P. Carlin. "Markov Chain Monte Carlo Methods for Computing Bayes Factors". In: *Journal of the American Statistical Association* 96.455 (2001), pp. 1122–1132. DOI: 10.1198/016214501753208780.

[63] P. Dellaportas, J. J. Forster, and I. Ntzoufras. "On Bayesian model and variable selection using MCMC". In: *Statistics and Computing* 12.1 (2002), pp. 27–36. DOI: 10.1023/A:1013164120801.

[64] A. O'Hagan and J. F. C. Kingman. "Curve Fitting and Optimal Design for Prediction". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 40.1 (1978), pp. 1–42.

[65] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. "Design and Analysis of Computer Experiments". In: *Statistical Science* 4.4 (1989), pp. 409–423. DOI: 10.1214/ss/1177012413.

[66] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. New York: Springer-Verlag, 1991. DOI: 10.1007/978-1-4612-3094-6.

[67] T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. New York: Springer, 2003. DOI: 10.1007/978-1-4757-3799-8.

[68] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts, USA: The MIT Press, 2006.

[69] O. P. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Scientific Computation. Dordrecht, Netherlands: Springer, 2010. DOI: 10.1007/978-90-481-3520-2.

[70] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton, New Jersey, USA: Princeton University Press, 2010.

[71] H. Stahl and V. Totik. *General Orthogonal Polynomials*. Encyclopedia of Mathematics and its Applications 43. Cambridge, UK: Cambridge University Press, 1992. DOI: 10.1017/CBO9780511759420.

[72] W. Gautschi. *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Oxford, UK: Oxford University Press, 2004.

[73] D. Jackson. *Fourier Series and Orthogonal Polynomials*. Dover Books on Mathematics. Mineola, New York, USA: Dover Publications, Inc., 2004. DOI: 10.5948/upo9781614440062.008.

[74] D. Xiu and G. E. Karniadakis. "The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations". In: *SIAM Journal on Scientific Computing* 24.2 (2002), pp. 619–644. DOI: 10.1137/S1064827501387826.

[75] J. A. S. Witteveen, S. Sarkar, and H. Bijl. "Modeling physical uncertainties in dynamic stall induced fluid–structure interaction of turbine blades using arbitrary polynomial chaos". In: *Computers & Structures* 85.11–14 (2007), pp. 866–878. DOI: 10.1016/j.compstruc.2007.01.004.

[76] G. E. P. Box and N. R. Draper. *Response Surfaces, Mixtures, and Ridge Analyses*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2007. DOI: 10.1002/0470072768.

[77] I. Babuška, R. Tempone, and G. E. Zouraris. "Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations". In: *SIAM Journal on Numerical Analysis* 42.2 (2004), pp. 800–825. DOI: 10.1137/S0036142902418680.

[78] D. Xiu and J. Shen. "Efficient stochastic Galerkin methods for random diffusion equations". In: *Journal of Computational Physics* 228.2 (2009), pp. 266–281. DOI: 10.1016/j.jcp.2008.09.008.

[79] D. Xiu and J. S. Hesthaven. "High-Order Collocation Methods for Differential Equations with Random Inputs". In: *SIAM Journal on Scientific Computing* 27.3 (2005), pp. 1118–1139. DOI: 10.1137/040615201.

[80] D. Xiu. "Efficient Collocational Approach for Parametric Uncertainty Analysis". In: *Communications in Computational Physics* 2.2 (2007), pp. 293–309.

[81] O. P. Le Maître, O. M. Knio, H. N. Najm, and R. G. Ghanem. "A Stochastic Projection Method for Fluid Flow: I. Basic Formulation". In: *Journal of Computational Physics* 173.2 (2001), pp. 481–511. DOI: 10.1006/jcph.2001.6889.

[82] O. P. Le Maître, M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio. "A Stochastic Projection Method for Fluid Flow: II. Random Process". In: *Journal of Computational Physics* 181.1 (2002), pp. 9–44. DOI: 10.1006/jcph.2002.7104.

[83] M. Berveiller, B. Sudret, and M. Lemaire. "Stochastic finite element: a non intrusive approach by regression". In: *European Journal of Computational Mechanics* 15.1–3 (2006), pp. 81–92. DOI: 10.3166/remn.15.81-92.

[84] G. Blatman and B. Sudret. "Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach". In: *Comptes Rendus Mécanique* 336.6 (2008), pp. 518–523. DOI: 10.1016/j.crme.2008.02.013.

[85] C. Giraud. *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability 139. Boca Raton, Florida, USA: CRC Press, 2015.

[86] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Monographs on Statistics and Applied Probability 143. Boca Raton, Florida, USA: CRC Press, 2015.

[87] G. Blatman and B. Sudret. "Adaptive sparse polynomial chaos expansion based on least angle regression". In: *Journal of Computational Physics* 230.6 (2011), pp. 2345–2367. DOI: 10.1016/j.jcp.2010.12.021.

[88] A. Doostan and H. Owhadi. "A non-adapted sparse approximation of PDEs with stochastic inputs". In: *Journal of Computational Physics* 230.8 (2011), pp. 3015–3034. DOI: 10.1016/j.jcp.2011.01.002.

[89] L. Yan, L. Guo, and D. Xiu. "Stochastic collocation algorithms using ℓ_1-minimization". In: *International Journal for Uncertainty Quantification* 2.3 (2012), pp. 279–293. DOI: 10.1615/Int.J.UncertaintyQuantification.2012003925.

[90] L. Mathelin and K. A. Gallivan. "A Compressed Sensing Approach for Partial Differential Equations with Random Input Data". In: *Communications in Computational Physics* 12.4 (2012), pp. 919–954. DOI: 10.4208/cicp.151110.090911a.

[91] K. Sargsyan, C. Safta, H. N. Najm, B. J. Debusschere, D. Ricciuto, and P. Thornton. "Dimensionality Reduction for Complex Models via Bayesian Compressive Sensing". In: *International Journal for Uncertainty Quantification* 4.1 (2014), pp. 63–93. DOI: 10.1615/Int.J.UncertaintyQuantification.2013006821.

[92] J. Ray, Z. Hou, M. Huang, K. Sargsyan, and L. Swiler. "Bayesian Calibration of the Community Land Model Using Surrogates". In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1 (2015), pp. 199–233. DOI: 10.1137/140957998.

[93] G. Karagiannis and G. Lin. "Selection of polynomial chaos bases via Bayesian model uncertainty methods with applications to sparse approximation of PDEs with stochastic inputs". In: *Journal of Computational Physics* 259 (2014), pp. 114–134. DOI: 10.1016/j.jcp.2013.11.016.

[94] G. Karagiannis, B. A. Konomi, and G. Lin. "A Bayesian mixed shrinkage prior procedure for spatial–stochastic basis selection and evaluation of gPC expansions: Applications to elliptic SPDEs". In: *Journal of Computational Physics* 284 (2015), pp. 528–546. DOI: 10.1016/j.jcp.2014.12.034.

[95] A. Cohen, M. A. Davenport, and D. Leviatan. "On the Stability and Accuracy of Least Squares Approximations". In: *Foundations of Computational Mathematics* 13.5 (2013), pp. 819–834. DOI: 10.1007/s10208-013-9142-3.

[96] G. Migliorati, F. Nobile, E. von Schwerin, and R. Tempone. "Analysis of Discrete L^2 Projection on Polynomial Spaces with Random Evaluations". In: *Foundations of Computational Mathematics* 14.3 (2014), pp. 419–456. DOI: 10.1007/s10208-013-9186-4.

[97] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone. "Discrete least squares polynomial approximation with random evaluations – application to parametric and stochastic elliptic PDEs". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 49.3 (2015), pp. 815–837. DOI: 10.1051/m2an/2014050.

[98] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. "Least angle regression". In: *The Annals of Statistics* 32.2 (2004), pp. 407–499. DOI: 10.1214/009053604000000067.

[99] T. Hesterberg, N. H. Choi, L. Meier, and C. Fraley. "Least angle and ℓ_1 penalized regression: A review". In: *Statistics Surveys* 2 (2008), pp. 61–93. DOI: 10.1214/08-SS035.

[100] S. Arlot and A. Celisse. "A survey of cross-validation procedures for model selection". In: *Statistics Surveys* 4 (2010), pp. 40–79. DOI: 10.1214/09-SS054.

[101] G. A. F. Seber and A. J. Lee. *Linear Regression Analysis*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2003. DOI: 10.1002/9780471722199.

[102] S. Efromovich. "Orthogonal series density estimation". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4 (2010), pp. 467–476. DOI: 10.1002/wics.97.

[103] M. Jiang and S. B. Provost. "Improved orthogonal polynomial density estimates". In: *Journal of Statistical Computation and Simulation* 81.11 (2011), pp. 1495–1516. DOI: 10.1080/00949655.2010.492781.

[104] R. A. Johnson. "An Asymptotic Expansion for Posterior Distributions". In: *The Annals of Mathematical Statistics* 38.6 (1967), pp. 1899–1906. DOI: 10.1214/aoms/1177698624.

[105] R. A. Johnson. "Asymptotic Expansions Associated with Posterior Distributions". In: *The Annals of Mathematical Statistics* 41.3 (1970), pp. 851–864. DOI: 10.1214/aoms/1177696963.

[106] R. C. Weng. "A Bayesian Edgeworth expansion by Stein's identity". In: *Bayesian Analysis* 5.4 (2010), pp. 741–763. DOI: 10.1214/10-BA526.

[107] R. C. Weng and C.-H. Hsu. "A Study of Expansions of Posterior Distributions". In: *Communications in Statistics - Theory and Methods* 42.2 (2013), pp. 346–364. DOI: 10.1080/03610926.2011.579701.

[108] H. R. B. Orlande, M. J. Colaço, and G. S. Dulikravich. "Approximation of the likelihood function in the Bayesian technique for the solution of inverse problems". In: *Inverse Problems in Science and Engineering* 16.6 (2008), pp. 677–692. DOI: 10.1080/17415970802231677.

[109] A. Dietzel and P. Reichert. "Bayesian inference of a lake water quality model by emulating its posterior density". In: *Water Resources Research* 50.10 (2014), pp. 7626–7647. DOI: 10.1002/2012WR013086.

[110] C. Soize and R. Ghanem. "Physical Systems with Random Uncertainties: Chaos Representations with Arbitrary Probability Measure". In: *SIAM Journal on Scientific Computing* 26.2 (2004), pp. 395–410. DOI: 10.1137/S1064827503424505.

[111] D. E. Barndorff-Nielsen and D. R. Cox. *Asymptotic Techniques for Use in Statistics*. Monographs on Statistics and Applied Probability 31. London, UK: Chapman & Hall/CRC, 1989.

[112] J. E. Kolassa. *Series Approximation Methods in Statistics*. 3rd ed. Lecture Notes in Statistics 88. New York: Springer, 2006. DOI: 10.1007/0-387-32227-2.

[113] C. G. Small. *Expansions and Asymptotics for Statistics*. Monographs on Statistics and Applied Probability 115. Boca Raton, Florida, USA: Chapman & Hall/CRC, 2010. DOI: 10.1201/9781420011029.

[114] S. Marelli and B. Sudret. "UQLab: A Framework for Uncertainty Quantification in Matlab". In: *2nd International Conference on Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management (ICVRAM-ISUMA 2014)*. Ed. by M. Beer, S.-K. Au, and J. W. Hall. Reston, Virginia, USA: American Society of Civil Engineers (ASCE), 2014. Chap. 257, pp. 2554–2563. DOI: 10.1061/9780784413609.257.

[115] S. Marelli and B. Sudret. *UQLab User Manual: Polynomial Chaos Expansions*. Version 0.9-104. Chair of Risk, Safety & Uncertainty Quantification, ETH Zürich. 2015.

Part III

Further work

Chapter 9

Bayesian inference as a random variable transformation

This chapter provides an introduction to and demonstration of Bayesian inference via transport maps. Bayesian updating is first recast as a random variable transformation and then solved as an optimization problem. The latter involves an information-theoretic optimality criterion. In particular, the relative entropy of the back-transformed posterior from the prior is minimized. After posing the problem that way, it becomes regularized in the framework of optimal transportation theory.

The transport map–based formulation was originally introduced in [1]. A nice overview of the optimal transportation of probability measures for purposes of Bayesian inference is given in [2]. Similar ideas had also emerged in the context of sequential data assimilation, an overview of which can be found in [3, 4]. Transformation-based inference joins the ranks of the Bayesian methods reviewed in Section 3.7. It can be seen as a special case of variational Bayesian inference in Section 3.7.2, where certain prior transformations constitute the parametric family of candidate distributions. Beyond that, it shares commonalities with spectral Bayesian inference as presented in Chapter 8.

A simple inverse heat conduction problem is used for demonstration purposes. Unknown thermal conductivities of a composite material are indirectly inferred from measurements of the temperature that are taken close to the boundary. The prior is transformed into the corresponding posterior distribution. Traditional Markov chain Monte Carlo sampling serves as the reference solution. Parts of this chapter were also presented at the International Symposium on Reliability of Engineering Systems that was held on October 15–17, 2015 in Hangzhou, China [5].

The tutorial on Bayesian inference as a random variable transformation is structured as follows. In Section 9.1 the optimal transportation from the prior to the posterior is investigated. In Section 9.2 a variational problem is formulated that allows for a numerical computation of the posterior. Subsequently, Section 9.3 covers practical issues such as the parametrization of the map and the regularization of its computation. In Section 9.4 the pros and cons of the approach are weighed and compared to spectral Bayesian inference that was devised previously. In Section 9.5 an inverse heat conduction problem is solved as an illustrative example of transformation-based Bayesian inference. Finally, Section 9.6 contains some concluding remarks.

9.1 Prior transformations

Given the prior distribution and the likelihood function, Bayes' rule characterizes the posterior as a conditional density. It is interesting to view the transition from the prior to the posterior density in terms of random variables instead. While that point of view is not directly suggested by Bayes's law, which only operates on probability densities, it indeed provides some useful intuition and leads to new recipes for computational Bayesian inference. This is investigated next.

Consider an injective and continuously differentiable map $T\colon \mathbb{R}^M \to \mathbb{R}^M$ that transforms a random variable $\boldsymbol{X} \sim \pi(\boldsymbol{x})$ distributed according to the prior distribution into a new random variable

$$\tilde{\boldsymbol{X}} = T(\boldsymbol{X}) \sim \pi_T(\tilde{\boldsymbol{x}}). \tag{9.1}$$

The map is assumed to be invertible and sufficiently well-behaved, such that one can write the transformed density of the random variable in Eq. (9.1) as

$$\pi_T(\tilde{\boldsymbol{x}}) = \pi(T^{-1}(\tilde{\boldsymbol{x}}))\,|\det J_{T^{-1}}(\tilde{\boldsymbol{x}})|. \tag{9.2}$$

Conversely, given that a random variable $\tilde{\boldsymbol{X}} \sim \pi(\tilde{\boldsymbol{x}}|\boldsymbol{y})$ is distributed according to the posterior, one may consider the back-transformation $\boldsymbol{X} = T^{-1}(\tilde{\boldsymbol{X}}) \sim \pi_{T^{-1}}(\boldsymbol{x}|\boldsymbol{y})$ with

$$\pi_{T^{-1}}(\boldsymbol{x}|\boldsymbol{y}) = \pi(T(\boldsymbol{x})|\boldsymbol{y})\,|\det J_T(\boldsymbol{x})| = \frac{\mathcal{L}(T(\boldsymbol{x}))\pi(T(\boldsymbol{x}))}{Z}\,|\det J_T(\boldsymbol{x})|. \tag{9.3}$$

One is interested in such maps T for which the transformed prior π_T in Eq. (9.2) equals the posterior and the back-transformed posterior $\pi_{T^{-1}}(\cdot|\boldsymbol{y})$ in Eq. (9.3) equates to the prior. This means that

$$\pi_T = \pi(\cdot|\boldsymbol{y}), \quad \pi_{T^{-1}}(\cdot|\boldsymbol{y}) = \pi. \tag{9.4}$$

In this sense, the prior and the posterior density transform into one another. Given the prior random vector $\boldsymbol{X} \sim \pi(\boldsymbol{x})$, the transformed random vector $\tilde{\boldsymbol{X}} \sim \pi(\tilde{\boldsymbol{x}}|\boldsymbol{y})$ follows the posterior. While Eq. (9.4) is formulated in terms of densities, it also establishes a *deterministic coupling* $(\boldsymbol{X}, T(\boldsymbol{X}))$ of the prior and the posterior probability measure [6, 7].

An illustration of this principle is given in Fig. 9.1, where both the prior and the posterior are Gaussian distributions. This is the same setup as already visualized in Fig. 3.1. Two linear transformations are shown that transform the prior into the posterior density. For more complex problems involving non-Gaussian distributions the transformations will be generally nonlinear, though.

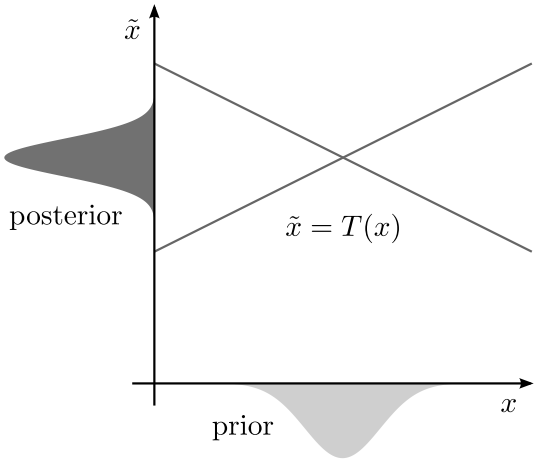


Figure 9.1: Prior transformation.

9.1.1 Couplings of Gaussians

Motivated by the illustrative example above, we study the case involving Gaussian distributions in greater depth. Consider two real-valued Gaussian random variables $X_1 \sim \mathcal{N}(x_1|\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(x_2|\mu_2, \sigma_2^2)$. Their respective distributions have means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. The first distribution transforms into the second one by

$$x_2 = T(x_1) = \mu_2 \pm \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1) \tag{9.5}$$

Two different linear couplings between the Gaussian distributions are described by Eq. (9.5) for which the variance transforms as $\sigma_2^2 = \sigma_1^2(\pm\sigma_2/\sigma_1)^2$. In the case that the random variables $X_1 \sim \mathcal{N}(x_1|\mu_1, \sigma_1^2) = \pi(x_1)$ and $X_2 \sim \mathcal{N}(x_2|\mu_2, \sigma_2^2) = \pi(x_2|\boldsymbol{y})$ are distributed according to the prior and the posterior of Gaussian shape, respectively, this is exactly the scenario encountered in Fig. 9.1.

Now consider two $\mathbb{R}^M$-valued random variables $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{X}_2 \sim \mathcal{N}(\boldsymbol{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. They have Gaussian distributions with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Let $\boldsymbol{\Sigma}^{1/2}$ denote the principle square root of a symmetric and positive-definite matrix $\boldsymbol{\Sigma}$ that is uniquely characterized by $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$. In a straightforward manner, a deterministic coupling between the multivariate normal distributions is established by

$$\boldsymbol{x}_2 = T(\boldsymbol{x}_1) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1). \tag{9.6}$$

Indeed one has $\boldsymbol{\Sigma}_2 = (\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1^{-1/2})\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1^{-1/2})^\top$. More generally, $\boldsymbol{\Sigma}_2 = (\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Phi}\boldsymbol{\Sigma}_1^{-1/2})\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Phi}\boldsymbol{\Sigma}_1^{-1/2})^\top$ for an arbitrary orthogonal matrix $\boldsymbol{\Phi}$ with $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \boldsymbol{I}$. Hence, the coupling in Eq. (9.6) is non-unique since any such $\boldsymbol{\Phi}$ defines an appropriate coupling by

$$\boldsymbol{x}_2 = T(\boldsymbol{x}_1) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Phi}\boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1). \tag{9.7}$$

The just-mentioned transforms define well-behaved couplings between Gaussian distributions that exhibit Jacobian formulas for the change of variables. More generally, non-continuous transformations may accomplish the same purpose, e.g. piecewise combinations of the linear transformations discussed, but cannot be written that nicely. In the following we exclusively concentrate on invertible and continuously differentiable maps. Beforehand, some introductory remarks on optimal transportation are given. This framework gives rise to important statements regarding the existence and uniqueness of deterministic couplings between general probability distributions, not only Gaussians.

9.1.2 Optimal transportation

As the preceding discussion revealed, a suitable map that transforms the prior into the posterior may not be unique. It may not even exist in the general case. In the framework of *optimal transport theory* [8, 9], however, one can establish certain existence and uniqueness results. A map T that satisfies $\pi_T = \pi(\cdot|\boldsymbol{y})$ is called a *transport map* in this context. Let a cost function $c\colon \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}^+$ represent the expense $c(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ of transporting a unit mass from $\boldsymbol{x}$ to $\tilde{\boldsymbol{x}}$. The total cost of the mapping $\tilde{\boldsymbol{x}} = T(\boldsymbol{x})$ is then

$$C(T) = \int_{\mathbb{R}^M} c(\boldsymbol{x}, T(\boldsymbol{x}))\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{9.8}$$

The *Monge problem* asks for finding a transport map that is optimal in that it minimizes the transportation cost in Eq. (9.8). As expressed in our density-oriented language and notation, this means to

$$\begin{aligned} &\text{minimize} && C(T), \\ &\text{subject to} && \pi_T = \pi(\cdot|\boldsymbol{y}). \end{aligned} \tag{9.9}$$

A solution to the problem in Eq. (9.9) is called an *optimal transport map*. Under relatively weak assumptions regarding the distributions involved and the cost function, one can ensure the existence and uniqueness of such an optimal map, e.g. for a quadratic cost function $c(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2^2$ such results are established by the *Brenier–McCann theorem* [10, 11]. Moreover, it states that the map is monotone. Under certain other cost considerations, the optimal map can be shown to coincide with the *Knothe–Rosenblatt rearrangement* [12, 13]. This means that the transport map has a triangular structure.

For the sake of completeness, it is remarked that the *Monge–Kantorovich problem* is a generalization of the problem discussed above. It allows for transport plans $\pi(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ and more general couplings $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ that admit the marginals $\boldsymbol{X} \sim \pi(\boldsymbol{x})$ and $\tilde{\boldsymbol{X}} \sim \pi(\tilde{\boldsymbol{x}}|\boldsymbol{y})$. The trivial coupling with $\pi(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \pi(\boldsymbol{x})\pi(\tilde{\boldsymbol{x}}|\boldsymbol{y})$ and the deterministic coupling with $\tilde{\boldsymbol{X}} = T(\boldsymbol{X})$ discussed above are two extreme cases. An optimal transference plan minimizes the total cost $C(T) = \iint_{\mathbb{R}^M \times \mathbb{R}^M} c(\boldsymbol{x}, \tilde{\boldsymbol{x}})\, \pi(\boldsymbol{x}, \tilde{\boldsymbol{x}})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\tilde{\boldsymbol{x}}$. Non-deterministic couplings are not permitted in Monge's original problem formulation which we actually focus on. However, for future research endeavors it may be useful to keep this possibility in mind.

In the Bayesian context, one is only interested in the transformation properties of the map and may thus safely disregard its cost and optimality. Indeed, here the map is merely an auxiliary construct with no associated transportation cost whatsoever. Hence, any arbitrary transport map is as good as any other. Yet, imposing an auxiliary cost function may still guide the practical computation of a transport map with a convenient structure, i.e. it regularizes the problem. Note that the quadratic cost function would favor maps close to the identity. This would for example suggest to pick the monotonically increasing transport map and scrap the decreasing one in Fig. 9.1 or Eq. (9.5). Other cost functions would promote other structures, e.g. triangular maps.

9.2 Variational formulation

For obtaining a transform-based representation of the posterior distribution, we have to solve Eq. (9.4) for an appropriate map or even crack the corresponding Monge problem in Eq. (9.9). Not to mention again the fact that the latter entails a completely pointless cost function, the two problems are extremely challenging. Unfortunately, as if traditional optimal transportation was not hard enough, the problem is aggravated by the usual intricacies of Bayesian inference, i.e. the target posterior distribution is only partially known through pointwise references to its unnormalized density.

In order to compute an approximate transport map despite all these difficulties, a variational formulation on the basis of Section 3.7.2 is devised. It relies on fundamental concepts in information theory [14, 15]. Instead of finding a map that exactly establishes a deterministic coupling between the prior and the posterior, one can

resort to a map that fits a certain information-theoretic optimality criterion. In this regard, the Kullback–Leibler (KL) divergence of the back-transformed posterior $\pi_{T^{-1}}(\cdot|\boldsymbol{y})$ from the prior π is considered

$$D_{\mathrm{KL}}(\pi\|\pi_{T^{-1}}(\cdot|\boldsymbol{y})) = \int_{\mathbb{R}^M} \log\left(\frac{\pi(\boldsymbol{x})}{\pi_{T^{-1}}(\boldsymbol{x}|\boldsymbol{y})}\right)\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \log Z - \mathcal{G}(T). \tag{9.10}$$

This shows an intriguing resemblance to Eq. (3.56). The divergence $D_{\mathrm{KL}}(\pi\|\pi_{T^{-1}}(\cdot|\boldsymbol{y}))$ is now the difference between the constant log-evidence $\log Z$ and

$$\begin{aligned}\mathcal{G}(T) &= \int_{\mathbb{R}^M} \log\left(\frac{\mathcal{L}(T(\boldsymbol{x}))\pi(T(\boldsymbol{x}))|\det J_T(\boldsymbol{x})|}{\pi(\boldsymbol{x})}\right)\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &= \int_{\mathbb{R}^M} \left(\log\mathcal{L}(T(\boldsymbol{x})) + \log\pi(T(\boldsymbol{x})) + \log|\det J_T(\boldsymbol{x})| - \log\pi(\boldsymbol{x})\right)\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{9.11}$$

In the best case one would have $\pi = \pi_{T^{-1}}(\cdot|\boldsymbol{y})$ such that $D_{\mathrm{KL}}(\pi\|\pi_{T^{-1}}(\cdot|\boldsymbol{y})) = 0$ and $\mathcal{G}(T) = \log Z$. More generally one has $D_{\mathrm{KL}}(\pi\|\pi_{T^{-1}}(\cdot|\boldsymbol{y})) \geq 0$, thus a variational lower bound of the evidence is established through $\log Z \geq \mathcal{G}(T)$ instead of the free energy in Eq. (3.57). Note that the differential Shannon entropy $H_{\mathrm{S}}(\pi) = -\int_{\mathbb{R}^M}\log(\pi(\boldsymbol{x}))\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ of the prior density emerges when one decomposes as $\mathcal{G}(T) = \int_{\mathbb{R}^M}\log(\mathcal{L}(T(\boldsymbol{x}))\pi(T(\boldsymbol{x}))|\det J_T(\boldsymbol{x})|)\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + H_{\mathrm{S}}(\pi)$.

Let us consider a class of possible transformations $\mathbb{T}$. Then one can find the member $T \in \mathbb{T}$ that, measured in terms of the relative entropy, best back-transforms the posterior into the prior $\pi_{T^{-1}}(\cdot|\boldsymbol{y})) \approx \pi$. Hence, the posterior is well approximated by the transformed prior $\pi_T \approx \pi(\cdot|\boldsymbol{y})$. Similar as in Eq. (3.59), minimizing $D_{\mathrm{KL}}(\pi\|\pi_{T^{-1}}(\cdot|\boldsymbol{y}))$ is equivalent to maximizing $\mathcal{G}(T)$. Thus

$$T = \operatorname*{arg\,min}_{T_\star\in\mathbb{T}} D_{\mathrm{KL}}(\pi\|\pi_{T_\star^{-1}}(\cdot|\boldsymbol{y})) \quad\Leftrightarrow\quad T = \operatorname*{arg\,max}_{T_\star\in\mathbb{T}} \mathcal{G}(T_\star). \tag{9.12}$$

Basically, this minimizes the information loss or entropy gain which comes along with replacing the prior with the back-transformed posterior. A stochastic program is posed in that a probabilistic expectation under the prior is extremized [16, 17].

As opposed to $D_{\mathrm{KL}}(\pi\|\pi_{T^{-1}}(\cdot|\boldsymbol{y}))$ in Eq. (9.10), the intractable model evidence has been eliminated from $\mathcal{G}(T)$ in Eq. (9.11). Instead of minimizing $D_{\mathrm{KL}}(\pi\|\pi_{T^{-1}}(\cdot|\boldsymbol{y}))$, one might hit on the alternative idea of doing so for $D_{\mathrm{KL}}(\pi_{T^{-1}}(\cdot|\boldsymbol{y})\|\pi) = \int_{\mathbb{R}^M}\log(\pi_{T^{-1}}(\boldsymbol{x}|\boldsymbol{y})/\pi(\boldsymbol{x}))\,\pi_{T^{-1}}(\boldsymbol{x}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}$. This would involve intractable expectations under the back-transformed posterior, however. Another idea is to minimize either of the distances $D_{\mathrm{KL}}(\pi(\cdot|\boldsymbol{y})\|\pi_T)$ or $D_{\mathrm{KL}}(\pi_T\|\pi(\cdot|\boldsymbol{y}))$ between the transformed prior and the posterior. In addition to the intractable expectation arising in the first objective, this is generally problematic because it would require to evaluate the inverse map T^{-1} and its Jacobian determinant $\det J_{T^{-1}}$. On the whole, the minimization of $D_{\mathrm{KL}}(\pi\|\pi_{T^{-1}}(\cdot|\boldsymbol{y}))$ in Eq. (9.12) is the only viable option.

9.3 Practical computation

After having reformulated Bayesian inference as a stochastic program, a few more ingredients are still necessary in order to render its practical solution feasible. This involves means to parametrize the map, regularize its computation and evaluate stochastic averages in the optimization routine. Herein we restrict the search for a transformation to a convenient and reasonably rich function space and deploy a Monte Carlo approximation of the optimization objective.

9.3.1 Map parametrization

We consider maps with a triangular structure. This form is motivated by the discussion about optimal transport where it emerges as a consequence of certain considerations regarding the transportation cost. For $i = 1,\dots,M$ each component $T_i(x_1,\dots,x_i)$ is a function of the first i variables only. Overall, such a triangular-like map is written as

$$T^{\triangle}(\boldsymbol{x}) = \begin{pmatrix} T_1(x_1)\\ T_2(x_1,x_2)\\ \vdots\\ T_M(x_1,x_2,\dots,x_M)\end{pmatrix}. \tag{9.13}$$

An apparent characteristic of this formulation is that it depends on the ordering of the variables involved. The Jacobian matrix of the vector function in Eq. (9.13) has the lower triangular structure

$$J_{T^\triangle} = \frac{\mathrm{d}T^\triangle}{\mathrm{d}\boldsymbol{x}} = \begin{pmatrix} \frac{\partial T_1}{\partial x_1} & 0 & \dots & 0 \\ \frac{\partial T_2}{\partial x_1} & \frac{\partial T_2}{\partial x_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial T_M}{\partial x_1} & \frac{\partial T_M}{\partial x_2} & \dots & \frac{\partial T_M}{\partial x_M} \end{pmatrix}. \tag{9.14}$$

As a direct consequence thereof, the determinant of the lower triangular matrix in Eq. (9.14) is simply the product of its diagonal terms

$$\det J_{T^\triangle} = \prod_{i=1}^{M} \frac{\partial T_i}{\partial x_i}. \tag{9.15}$$

That the Jacobian determinant can be easily determined through Eq. (9.15) is of great help. It simplifies the evaluation of objective function in Eq. (9.11) for the optimization problem in Eq. (9.12).

After specifying the structure of the random variable transformation, we have to represent its individual components in some way. Multivariate polynomials up to a certain degree are envisaged for that purpose. They provide a convenient basis for representing smooth functions which is both flexible and interpretable. Given that we have a candidate set of polynomials $\{\Psi_{\boldsymbol{\alpha}_i}(x_1, \dots, x_i)\}_{\boldsymbol{\alpha}_i \in \mathcal{A}_{i,p_i}}$ for all $i = 1, \dots, M$, the components of the map in Eq. (9.13) can be represented as a superposition

$$T_i(x_1, \dots, x_i) = \sum_{\boldsymbol{\alpha}_i \in \mathcal{A}_{i,p_i}} a_{\boldsymbol{\alpha}_i} \Psi_{\boldsymbol{\alpha}_i}(x_1, \dots, x_i). \tag{9.16}$$

As usual, multi-indices $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \dots, \alpha_{i,i}) \in \mathbb{N}^i$ are introduced in order to bookkeep and restrict the number of terms in Eq. (9.16). The total polynomial degrees $\|\boldsymbol{\alpha}_i\|_1 = \sum_{j=1}^{i} |\alpha_{i,j}| \leq p_i$ are limited to i-dependent maxima $p_i \in \mathbb{N}$. Only terms whose multi-index satisfies $\boldsymbol{\alpha}_i \in \mathcal{A}_{i,p_i} = \left\{\boldsymbol{\beta}_i \in \mathbb{N}^i \colon \|\boldsymbol{\beta}_i\|_1 \leq p_i\right\}$ are then kept. Let us now fix the polynomial degrees $p_i = p$ to an identical upper limit $p \in \mathbb{N}$ for all components of the map. Moreover, letting $\boldsymbol{a} \in \mathbb{R}^P$ with $P \in \mathbb{N}_{>0}$ denote the totality of the coefficients for all expansions, one can write

$$T^\triangle_{\boldsymbol{a}}(\boldsymbol{x}) = \begin{pmatrix} \sum\limits_{\boldsymbol{\alpha}_1 \in \mathcal{A}_{1,p}} a_{\boldsymbol{\alpha}_1} \Psi_{\boldsymbol{\alpha}_1}(x_1) \\ \sum\limits_{\boldsymbol{\alpha}_2 \in \mathcal{A}_{2,p}} a_{\boldsymbol{\alpha}_2} \Psi_{\boldsymbol{\alpha}_2}(x_1, x_2) \\ \vdots \\ \sum\limits_{\boldsymbol{\alpha}_M \in \mathcal{A}_{M,p}} a_{\boldsymbol{\alpha}_M} \Psi_{\boldsymbol{\alpha}_M}(x_1, x_2, \dots, x_M) \end{pmatrix}. \tag{9.17}$$

The number of terms in Eq. (9.17) grows fast with increasing dimensionality M and expansion order p. Based on Eq. (2.24) the total number of terms P is given as

$$P = \sum_{i=1}^{M} \binom{i+p}{p} = \sum_{i=1}^{M} \frac{(i+p)!}{i!\,p!}. \tag{9.18}$$

The triangular formulation of the map with polynomial components gives rise to one free algorithmic parameter that has to be set, namely the maximal polynomial degree p.

While there is no stringent necessity for choosing a family of polynomials that is orthogonal with respect to the prior weight function, it certainly is appealing to do so. This may require a transformation to standardized variables, which is tantamount to the introduction of a secondary measure that transforms into both the prior and the posterior. The random vector in Eq. (9.1) can then be seen as a triangular type of polynomial chaos expansion $\tilde{\boldsymbol{X}} = T^\triangle_{\boldsymbol{a}}(\boldsymbol{X})$ in terms of the random variables $\boldsymbol{X} \sim \pi(\boldsymbol{x})$ that are distributed according to the prior. Notwithstanding that this facilitates the interpretation of the coefficients of the parametrized map, we do not aim at leveraging the Hilbert space theory from Section 2.3 per se. Rather we just need any class of transformations that is sufficiently flexible, i.e. in order to contain such members that well couple the prior and the posterior, and reasonably restrictive at the same time, i.e. so as to facilitate the process of finding a good transform.

Analogous to Eq. (2.26), given that the prior is a product measure with a density $\pi(\boldsymbol{x}) = \pi_1(x_1)\dots\pi_M(x_M)$ and that the basis functions in Eq. (9.16) are normalized, one can write the means and variances of the random variables $T_i(X_1,\dots,X_i)$ for $i = 1,\dots,M$ as

$$\mathbb{E}[T_i(X_1,\dots,X_i)] = a_{\mathbf{0}_i}, \quad \mathrm{Var}[T_i(X_1,\dots,X_i)] = \sum_{\boldsymbol{\alpha}_i \in \mathcal{A}_{i,p}\setminus\{\mathbf{0}_i\}} a^2_{\boldsymbol{\alpha}_i}. \tag{9.19}$$

The zero vector of $\mathbb{R}^i$ is here denoted as $\mathbf{0}_i$. For $i,j = 1,\dots,M$ with $j > i$ one can similarly write the covariance between any two different random variables $T_i(X_1,\dots,X_i)$ and $T_j(X_1,\dots,X_j)$ as

$$\mathrm{Cov}[T_i(X_1,\dots,X_i), T_j(X_1,\dots,X_j)] = \sum_{\boldsymbol{\alpha}_i \in \mathcal{A}_{i,p}\setminus\{\mathbf{0}_i\}} a_{\boldsymbol{\alpha}_i} a_{(\boldsymbol{\alpha}_i,\mathbf{0}_{j-i})}. \tag{9.20}$$

Here, $(\boldsymbol{\alpha}_i,\mathbf{0}_{j-i}) = (\alpha_{i,1},\dots,\alpha_{i,i},0,\dots,0) \in \mathcal{A}_{j,p}$ denotes the concatenation of the multi-index $\boldsymbol{\alpha}_i \in \mathbb{R}^i$ and the zero element $\mathbf{0}_{j-i} = (0,\dots,0) \in \mathbb{R}^{j-i}$.

Provided that an inferential map perfectly couples the prior and the posterior, Eqs. (9.19) and (9.20) immediately provide the first posterior moments. Conditional on the data, the posterior means and covariances are simply given as $\mathbb{E}[X_i|\boldsymbol{y}] = \mathbb{E}[T_i(X_1,\dots,X_i)]$ and $\mathrm{Cov}[X_i,X_j|\boldsymbol{y}] = \mathrm{Cov}[T_i(X_1,\dots,X_i), T_j(X_1,\dots,X_j)]$, respectively. In case a transport map only establishes an imperfect coupling, these relations may still serve as posterior approximations.

9.3.2 Sample average approximation

For evaluating the objective function, we avail ourselves of Monte Carlo (MC) simulation. In principle one can imagine two different modalities of random sampling–based stochastic optimization, i.e. one can either use a fixed sample or resample for every computation of the objective function. The former approach is known as the *sample average approximation* (SAA) in stochastic programming [18, 19]. After the selection of the initial sample, the objective function is deterministic in that it always returns the same output value for the same values of the optimization parameters. The SAA can be readily implemented with any appropriate deterministic or stochastic optimizer. Its results are similarly straightforward to interpret. The latter approach features a stochastic approximation of the objective function in that it attains randomly varying outputs for the same inputs. While it can be implemented as easily as the SAA, its results are more problematic to interpret. For example, the algorithm could randomly but prematurely terminate due to obeying a stopping rule.

For these reasons, we use the SAA from now on. Let $(\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(K)})$ be a representative sample from the prior distribution of size $K \in \mathbb{N}_{>0}$, e.g. randomly and independently sampled. An MC approximation of the utility function in Eq. (9.11) is then given as

$$\hat{\mathcal{G}}(\boldsymbol{a}) = \frac{1}{K}\sum_{k=1}^{K}\left(\log\mathcal{L}(T^{\triangle}_{\boldsymbol{a}}(\boldsymbol{x}^{(k)})) + \log\pi(T^{\triangle}_{\boldsymbol{a}}(\boldsymbol{x}^{(k)})) + \log|\det J_{T^{\triangle}_{\boldsymbol{a}}}(\boldsymbol{x}^{(k)})| - \log\pi(\boldsymbol{x}^{(k)})\right). \tag{9.21}$$

The sample remains fixed once it has been selected. Given the *sample average function* in Eq. (9.21) and the parametrized transformation in Eq. (9.17), the maximization in Eq. (9.12) can be approximately stated as

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}^\star \in \mathbb{R}^P} \hat{\mathcal{G}}(\boldsymbol{a}^\star). \tag{9.22}$$

Two remaining choices have to be made for a practical implementation of the SAA method. A convenient optimizer has to be decided on and the MC sample size K has to be fixed.

9.3.3 Assessing convergence

Finally, an independent convergence criterion is desired. This refers to the iterations of the optimization algorithm, where the maximum achievable utility is unknown, but also to an additional outer loop that iterates over the maximal polynomial degree. Taking the logarithm of the back-transform in Eq. (9.3) subject to Eq. (9.4) and solving for the log-evidence leads to

$$\log Z = \log\mathcal{L}(T(\boldsymbol{x})) + \log\pi(T(\boldsymbol{x})) + \log|\det J_T(\boldsymbol{x})| - \log\pi(\boldsymbol{x}). \tag{9.23}$$

This characterizes the perfect coupling. However, the exact equality may not be achieved once the triangular map with polynomial components in Eq. (9.17) is used. In this case one can still utilize the right hand side of Eq. (9.23) in defining

$$\Lambda_{\boldsymbol{a}}(\boldsymbol{x}) = \log\mathcal{L}(T^{\triangle}_{\boldsymbol{a}}(\boldsymbol{x})) + \log\pi(T^{\triangle}_{\boldsymbol{a}}(\boldsymbol{x})) + \log|\det J_{T^{\triangle}_{\boldsymbol{a}}}(\boldsymbol{x})| - \log\pi(\boldsymbol{x}). \tag{9.24}$$

Note that the prior expectation of Eq. (9.24) is related to Eq. (9.11) by $\mathbb{E}[\Lambda_{\boldsymbol{a}}(\boldsymbol{X})] = \mathcal{G}(T^{\triangle}_{\boldsymbol{a}})$. In case that $T^{\triangle}_{\boldsymbol{a}}$ would establish a perfect coupling, one would have $\Lambda_{\boldsymbol{a}}(\boldsymbol{x}) = \log Z$ for all $\boldsymbol{x} \in \mathbb{R}^M$. This implies that $\mathbb{E}[\Lambda_{\boldsymbol{a}}(\boldsymbol{X})] = \log Z$ and $\mathrm{Var}[\Lambda_{\boldsymbol{a}}(\boldsymbol{X})] = 0$. The following alternative to the optimization problem in Eq. (9.12) and its discretization in Eq. (9.22) is thus suggested

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}^\star \in \mathbb{R}^P}{\arg\min}\, \mathrm{Var}[\Lambda_{\boldsymbol{a}^\star}(\boldsymbol{X})]. \tag{9.25}$$

Since Eq. (9.25) does not admit a straightforward interpretation, e.g. with regard to the minimization of a divergence measure, we do not solve that problem directly. Nevertheless, we can monitor $\mathrm{Var}[\Lambda_{\hat{\boldsymbol{a}}}(\boldsymbol{X})] \to 0$ or its MC sample approximation as a convergence indicator during the solution of Eq. (9.22). The selection of the maximal polynomial degree can be finally based on this criterion.

9.4 Comparison to SLEs

It is interesting to compare the variational transport formulation of Bayesian inference in this chapter to the spectral likelihood expansion (SLE) picture promoted in Chapter 8. Both approaches specify the joint posterior density in a flexible and nonparametric fashion. Of course, the representations of the posterior do actually feature parameters. Their number is, however, not dependent on a classical underlying probability model with parametric families of distributions. Instead, a wide range of non-classical densities can be represented, e.g. all posterior probability densities arising from a likelihood function which is mean-square integrable with respect to the prior distribution in the case of SLEs.

In this chapter, the posterior has been represented as a transformation of the prior in Eqs. (9.2) and (9.4), i.e. the prior density composed with the back-transformation times the corresponding Jacobian determinant. Based on an SLE, the posterior density had been represented in Eqs. (8.39) and (8.41) as the product of the prior density and a linear combination of multivariate orthogonal polynomials. While the SLE-based posterior density is immediately evaluable as a function of the unknowns, the evaluation of the transformation-based posterior density calls for the inverse map and its Jacobian. In the following, these representations based on orthogonal series expansions and probability density transforms are compared with each other in some more detail.

Both approaches to Bayesian inference are based on orthogonal polynomials, either for the expansion of probability densities or for the transformation of random variables. This is not imperative but very convenient. On the one hand, the scalar-valued likelihood in Eq. (8.35) is spectrally expanded in an orthonormal function space basis in Eqs. (8.36) and (8.37). This allows one to fully capitalize on Hilbert space theory. On the other hand, orthogonal polynomials provide a sufficiently adjustable representation of the transformation as the vector-valued function in Eq. (9.17). Properly done, a polynomial chaos expansion of the random vector in Eq. (9.1) distributed according to the posterior arises.

As for the spectral Bayesian approach, the SLE admits a statistically meaningful interpretation of the expansion coefficients. They are directly related to characteristics of the posterior distribution such as the model evidence in Eq. (8.40), the first posterior moments in Eqs. (8.48) to (8.50) and more general quantity of interest–posterior expectations in Eq. (8.43). Moreover, the posterior marginals emerge as sub-SLEs in Eqs. (8.45) and (8.47). In transport map inference, the coefficients of the parametrized transformation are also interpretable, albeit not so conveniently. The fact of the matter is that the model evidence is only accessible through an approximate lower bound. In order to estimate the first posterior moments one can use the relations in Eqs. (9.19) and (9.20). Nonetheless, for more general posterior expectation values one has to resort to a sampling procedure with independent draws. This is a limitation of the transportation approach and yet an advantage over the SLE method. While SLEs enable the evaluation of the normalized posterior density, they do not allow one to sample from the posterior distribution, at least not straightforwardly.

In spectral Bayesian inference the posterior is expanded as kind of a perturbation series about the prior as the reference density. This may require high-order expansion terms in case the posterior is significantly different from the prior. The baseline density change performed in Eqs. (8.51) and (8.52) clears that obstacle. It allows one to express the posterior in Eq. (8.55) as a correction to an auxiliary expansion density. In transport map inference the problem of higher-order terms is alleviated right from the start. That is because random variable transformations are highly effective in moving from one distribution to a completely different one. As exemplified through Eqs. (9.5) and (9.7), any arbitrary two Gaussian distributions can be linearly transformed into one another, no matter what the location and dispersion parameters. Moderately nonlinear maps may suffice for transporting between more general distributions.

So far we compared how spectral and transformation-based variational Bayesian inference characterize the posterior density function. Now we proceed with the practical computations. In order to compute an SLE one has to solve a linear stochastic program of the form as in Eq. (8.29). The linear least squares minimization in

Eq. (8.31) is a discretized variant of that problem. It has the appealingly simple ordinary least squares solution in Eq. (8.32). The leave-one-out error in Eq. (8.34) facilitates the selection of the sample size and expansion order. For the computation of a suitable coupling between the prior and posterior one has to solve the nonlinear stochastic optimization problem in Eq. (9.12) through the discretization in Eq. (9.22). In principle, this is a more complex problem than linear least squares. An independent convergence criterion is established by the variance of Eq. (9.24) under the prior distribution.

While both of the discussed approaches establish novel alternatives to traditional techniques in computational Bayesian inference, it is remarked that they have their own characteristic flaws. The emergence of negative values in the approximations of the likelihood function and posterior density surely is a weakness of the SLE method. A shortcoming of inferential transportation is that the formulation actually presupposes the invertibility of the candidate maps. However, this is likely violated by the polynomial representation.

9.5 Numerical experiment

Previously we discussed inferential transportation as firstly developed in [1, 2]. Many more interesting ideas were proposed in the original literature, e.g. the composition of transport maps. As a matter of fact, one does not need to transform the prior into the posterior in one step, which could possibly require a highly nonlinear map. Instead, one might construct a sequence of maps that establish an appropriate coupling in a step-by-step manner, i.e. one would only have to transport between similar interim distributions that progressively evolve into the target posterior. This lowers the necessary degree of nonlinearity. We do not investigate these ideas here, but focus on the one-step formulation.

A numerical experiment is conducted in order to demonstrate and study the transformation method. Its applicability for probabilistic parameter estimation is confirmed and its features and shortcomings are identified. For these purposes, an inverse heat conduction problem (IHCP) [20, 21] is solved. The setup is similar to the demonstration example used in the previous chapter. After calling the problem to mind, a numerical demonstration of transformation-based Bayesian inference is given. Markov chain Monte Carlo (MCMC) posterior sampling is employed as a reference and benchmarking solution.

9.5.1 Problem setup

We investigate a thermodynamic system with heat conduction in steady state, after a sufficiently long relaxation time has elapsed. The stationary heat equation $\nabla \cdot (\kappa \nabla \tilde{T}) = 0$ then governs the macroscopic diffusion of heat. Let $\tilde{T}(\boldsymbol{r})$ and $\kappa(\boldsymbol{r})$ denote the fields of temperature and thermal conductivity, respectively. They depend on the spatial coordinates $\boldsymbol{r} = (r_1, r_2)^\top$. Two space dimensions are considered.

A composite material with inclusions is the system under study. An illustration of the system and its geometry is provided in Fig. 9.2. The "top" of the domain is held at a fixed temperature $\tilde{T}_1$, which establishes a first-type boundary condition. At the "bottom" the heat flux $q_2 = -\kappa_0 \, \partial \tilde{T} / \partial r_2$ through the boundary is prescribed, which imposes a second-type condition. The "left" and "right" hand side are perfectly insulated such that there is no heat fluxing across the surfaces. In Table 9.1 the numeric values of the physical parameters including the boundary conditions are listed.

While the thermal conductivity κ_0 of the background matrix is considered well-known, the $M = 4$ conductivities $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \kappa_3, \kappa_4)^\top$ of the material inclusions are the unknown parameters. Their statistical inference is the goal of the Bayesian IHCP. A number of $N = 16$ data points of the temperature $\tilde{\boldsymbol{y}} = (\tilde{T}(\boldsymbol{r}_1), \dots, \tilde{T}(\boldsymbol{r}_N))^\top$ at the sensor locations $(\boldsymbol{r}_1, \dots, \boldsymbol{r}_N)$ is measured and analyzed to that end. The forward model $\mathcal{M} \colon \boldsymbol{\kappa} \mapsto \tilde{\boldsymbol{y}}$ arises from solving the boundary value problem corresponding to the partial differential equation for the measurable temperature as function of the unknowns. Here, the finite element method is used together with an interpolation of the nodal values to the sensor locations. A polynomial chaos expansion–based surrogate of the forward model is subsequently used in the Bayesian analysis.

The data $\boldsymbol{y} = (T(\boldsymbol{r}_1), \dots, T(\boldsymbol{r}_N))^\top$ comprise the observations of the temperature field at the measurement locations. They are thought of as $\boldsymbol{y} = \tilde{\boldsymbol{T}} + \boldsymbol{\varepsilon}$, i.e. as the model predictions $\tilde{\boldsymbol{T}} = \mathcal{M}(\boldsymbol{\kappa})$ corrupted with noise $\boldsymbol{\varepsilon}$. A multivariate Gaussian distribution $\pi(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon} | \boldsymbol{0}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ represents the measurement noise. The standard deviation σ quantifies the noise level. It is specified as $\sigma = 0.25\,\mathrm{K}$ in the numerical experiment. This establishes the data model $\boldsymbol{Y} | \boldsymbol{\kappa} \sim \mathcal{N}(\boldsymbol{y} | \mathcal{M}(\boldsymbol{\kappa}), \sigma^2 \boldsymbol{I})$ and the corresponding likelihood function $\mathcal{L}(\boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{y} | \mathcal{M}(\boldsymbol{\kappa}), \sigma^2 \boldsymbol{I})$. Herein we use pseudo-data that has been simulated according to the probability model just described.

We select a multivariate Gaussian prior distribution $\pi(\boldsymbol{\kappa}) = \prod_{i=1}^{4} \pi(\kappa_i)$ with independent marginals $\pi(\kappa_i) = \mathcal{N}(\kappa_i | \mu_0, \sigma_0^2)$. The mean and standard deviations are respectively specified as $\mu_0 = 30\,\mathrm{W/m/K}$ and $\sigma_0 = 5\,\mathrm{W/m/K}$. Even though this prior setup allows for negative thermal conductivities in principle, they are six

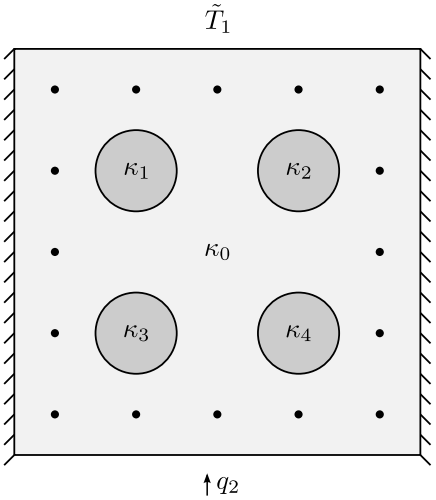


Figure 9.2: Heat conduction setup.

Table 9.1: Numeric parameter values.

κ_0	κ_1	κ_2	κ_3	κ_4	$\tilde{T}_1$	q_2
	[W/m/K]				[K]	[W/m^2]
15	20	27	33	40	200	2000

standard deviations far away from the mean. The prior probability mass assigned to those unphysical values is thus negligibly small. After the IHCP setup has been finally completed, for the density of the posterior distribution one has $\pi(\boldsymbol{\kappa}|\boldsymbol{y}) = Z^{-1}\mathcal{L}(\boldsymbol{\kappa})\pi(\boldsymbol{\kappa})$.

9.5.2 Algorithmic implementation

As already mentioned, many techniques from nonlinear programming [22–24] are now readily available for computing the posterior distribution by transforming the prior appropriately. We employ a quasi-Newton method where, in general black-box fashion, the necessary derivatives are smartly calculated by finite-differencing. In particular, we use the *Broyden–Fletcher–Goldfarb–Shanno* (BFGS) *algorithm* [25–29], an iterative quasi-Newton method for unconstrained nonlinear optimization. It is arguably one of the most widespread algorithms for this type of programming problems.

Several other local and global optimizers were tried out. This involves derivative-free optimization algorithms such as the Nelder–Mead method and pattern search. Moreover, this includes techniques from evolutionary computing such as a genetic algorithm, particle swarm optimization and the covariance matrix adaption evolution strategy. A typology of these diverse approaches eludes the scope of this chapter. The interested reader is redirected to [30, 31] for comprehensive overviews. As compared to BFGS, the performance of these techniques turns out to be rather mediocre. Hence, the BFGS algorithm is selected.

Multivariate normalized Hermite polynomials in conjunction with a standardizing parameter transform are used to build the triangular map. In order to avoid a too complex notation, the linear reparametrization remains implicit in the discussion of the inferential coupling. The sample average utility function is maximized as in Eq. (9.22) in a series of preliminary runs, where different values of K in Eq. (9.21) and p in Eq. (9.17) are tested. Apart from the comparison with the results from MCMC simulation, which hardly establishes a stand-alone solution, the only criterion at hand for deciding on those parameters is that the variance of Eq. (9.24) should be minimal as in Eq. (9.25). On this basis, we eventually choose a rather high sample size $K = 10^5$ for the SAA and a very low maximal polynomial degree $p = 2$. According to Eq. (9.18) we then have to find $P = 34$ unknown coefficients $\boldsymbol{a}$ of the transport map $T^{\triangle}_{\boldsymbol{a}}(\boldsymbol{x})$.

The BFGS algorithm is initialized at the identity map. It starts from the values $\boldsymbol{a}_0$ for which $T^{\triangle}_{\boldsymbol{a}_0}(\boldsymbol{x}) = \boldsymbol{x}$ is the identity map and then, over the course of the optimization, gradually transforms the prior into the posterior. It is stopped after $I = 51$ iterations and roughly three hours of program runtime, when the first-order optimality criterion $\|\nabla\hat{\mathcal{G}}(\boldsymbol{a}_I)\|_\infty \leq 10^{-6}$ is fulfilled. The final estimates of the coefficients that determine the inferential map $T^{\triangle}_{\hat{\boldsymbol{a}}}(\boldsymbol{x})$ are given as $\hat{\boldsymbol{a}} = \boldsymbol{a}_I$. We obtain $\hat{\mathcal{G}}(\hat{\boldsymbol{a}}) = -9.22$ and $\exp(\hat{\mathcal{G}}(\hat{\boldsymbol{a}})) = 9.86 \times 10^{-5}$ for the maximal value of the utility and its exponential. In Table 9.2 the quantities $\exp(\hat{\mathcal{G}}(\boldsymbol{a}_\iota))$ and $\mathrm{Var}[\Lambda_{\boldsymbol{a}_\iota}(\boldsymbol{X})]$ are listed for the intermediate values of the coefficients $\boldsymbol{a}_\iota$ that are obtained after every tenth iteration $\iota \in \{0, 1, 10, 20, 30, 40, 50\}$. While the evidence-related quantity $\exp(\hat{\mathcal{G}}(\boldsymbol{a}_\iota))$ is maximized, $\mathrm{Var}[\Lambda_{\boldsymbol{a}_\iota}(\boldsymbol{X})]$ is a separate convergence indicator. It attains $\mathrm{Var}[\Lambda_{\hat{\boldsymbol{a}}}(\boldsymbol{X})] = 0.73$ in the last iteration.

It is interesting to monitor the convergence of the coefficients over the BFGS iterations. In Fig. 9.3 the components of the coefficient vector $\boldsymbol{a}_\iota$ are shown for $\iota = 0, \ldots, 50$. The constant, linear and quadratic terms can be distinguished by their color. As it can be seen, the coefficients have almost reached their final values after

Table 9.2: BFGS optimization.

Iteration no. ι	0	1	10	20	30	40	50
$\exp(\hat{\mathcal{G}}(\boldsymbol{a}_\iota))$	4.23×10^{-86}	1.48×10^{-58}	3.25×10^{-5}	9.63×10^{-5}	9.86×10^{-5}	9.86×10^{-5}	9.86×10^{-5}
$\mathrm{Var}[\Lambda_{\boldsymbol{a}_\iota}(\boldsymbol{X})]$	2.44×10^{4}	1.08×10^{4}	6.71×10^{0}	8.52×10^{-1}	7.29×10^{-1}	7.31×10^{-1}	7.31×10^{-1}

about ten to twenty iterations and hardly change thereafter. Accordingly, we could have stopped the algorithm at this point already. Notice that the constant terms, which establish the posterior mean vector, start from the prior means and then approach the true values of the unknown thermal conductivities. The remaining coefficients of the linear and quadratic terms concentrate around lower values. They determine further characteristics of the posterior distribution such as the variances and correlations.

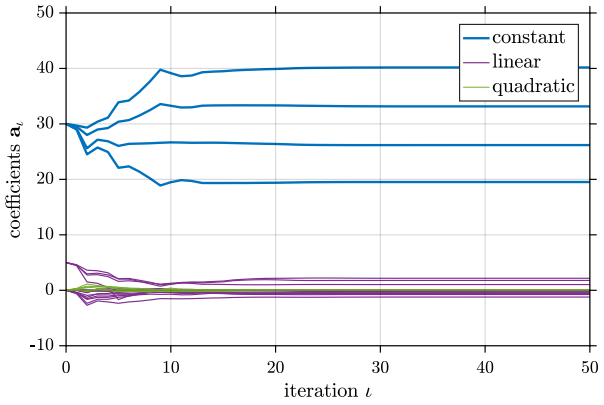


Figure 9.3: Converging coefficients.

The means and $\mathbb{E}[\kappa_i|\boldsymbol{y}]$ standard deviations $\mathrm{Std}[\kappa_i|\boldsymbol{y}] = \mathrm{Var}[\kappa_i|\boldsymbol{y}]^{1/2}$ of the posterior distribution for $i = 1, 2, 3, 4$ over the first twenty BFGS iterations are listed in Table 9.3. After every second algorithm iteration, the current values of the optimization parameters are used in the calculation of the two first posterior moments according to Eqs. (9.19) and (9.20). As it is observed, the expected values of the transformed random variables evolve from the prior into the posterior means. At the same time, the standard deviations expectedly decrease from the prior to the posterior level. The final estimates of the discussed posterior characteristics are further analyzed below.

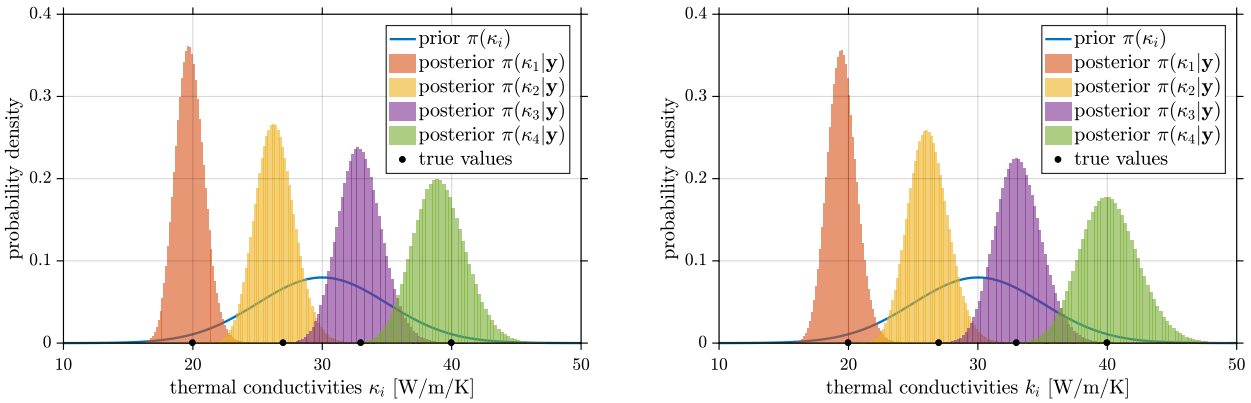
Table 9.3: Converging moments.

Iteration no. ι		0	2	4	6	8	10	12	14	16	18	20
$\mathbb{E}[\kappa_1\|\boldsymbol{y}]$	[W/m/K]	30	24.52	24.93	22.35	20.19	19.49	19.71	19.34	19.35	19.36	19.39
$\mathbb{E}[\kappa_2\|\boldsymbol{y}]$	[W/m/K]	30	25.61	26.87	26.41	26.51	26.67	26.59	26.61	26.55	26.46	26.38
$\mathbb{E}[\kappa_3\|\boldsymbol{y}]$	[W/m/K]	30	28.00	29.25	30.68	32.48	33.27	32.99	33.33	33.35	33.35	33.34
$\mathbb{E}[\kappa_4\|\boldsymbol{y}]$	[W/m/K]	30	29.32	31.13	34.22	37.64	39.15	38.72	39.42	39.64	39.79	39.90
$\mathrm{Std}[\kappa_1\|\boldsymbol{y}]$	[W/m/K]	5	3.38	3.05	2.33	1.53	1.23	1.33	1.16	1.11	1.10	1.10
$\mathrm{Std}[\kappa_2\|\boldsymbol{y}]$	[W/m/K]	5	4.31	2.88	2.60	1.97	1.80	1.82	1.59	1.58	1.57	1.57
$\mathrm{Std}[\kappa_3\|\boldsymbol{y}]$	[W/m/K]	5	3.65	3.21	2.05	1.17	1.08	1.41	1.44	1.74	1.90	1.89
$\mathrm{Std}[\kappa_4\|\boldsymbol{y}]$	[W/m/K]	5	3.43	2.93	2.04	1.50	1.38	1.59	1.66	1.94	2.15	2.22

9.5.3 Posterior distribution

After an appropriate random variable transformation has been found, the posterior distribution can be analyzed in view of its marginals, statistical moments and the like. The most general way of doing so is to sample the posterior. To that end one draws independent samples from the prior and applies the computed transformation to each of them individually. Independent samples from the posterior result from this procedure. They can be subsequently analyzed in order to visualize the posterior marginals or to compute conditional expectation values.

For the analysis of the marginal distributions, a prior sample of the size $L = 10^7$ is used. The same total number of samples is also computed by means of MCMC with thirty parallel chains. A comparison of the four posterior marginals is found in Fig. 9.4. In Fig. 9.4(a) histograms of the obtained MCMC sample are depicted. Directly besides in Fig. 9.4(b) histograms of the map-based posterior marginals are plotted. As far as one can tell by visual inspection, the marginals obtained from both methods are nearly identical. Only a minor deviation shows up in the fourth marginal. This means that the posterior marginals are captured very well with a low-degree triangular transformation.



(a) MCMC. (b) Mapping.

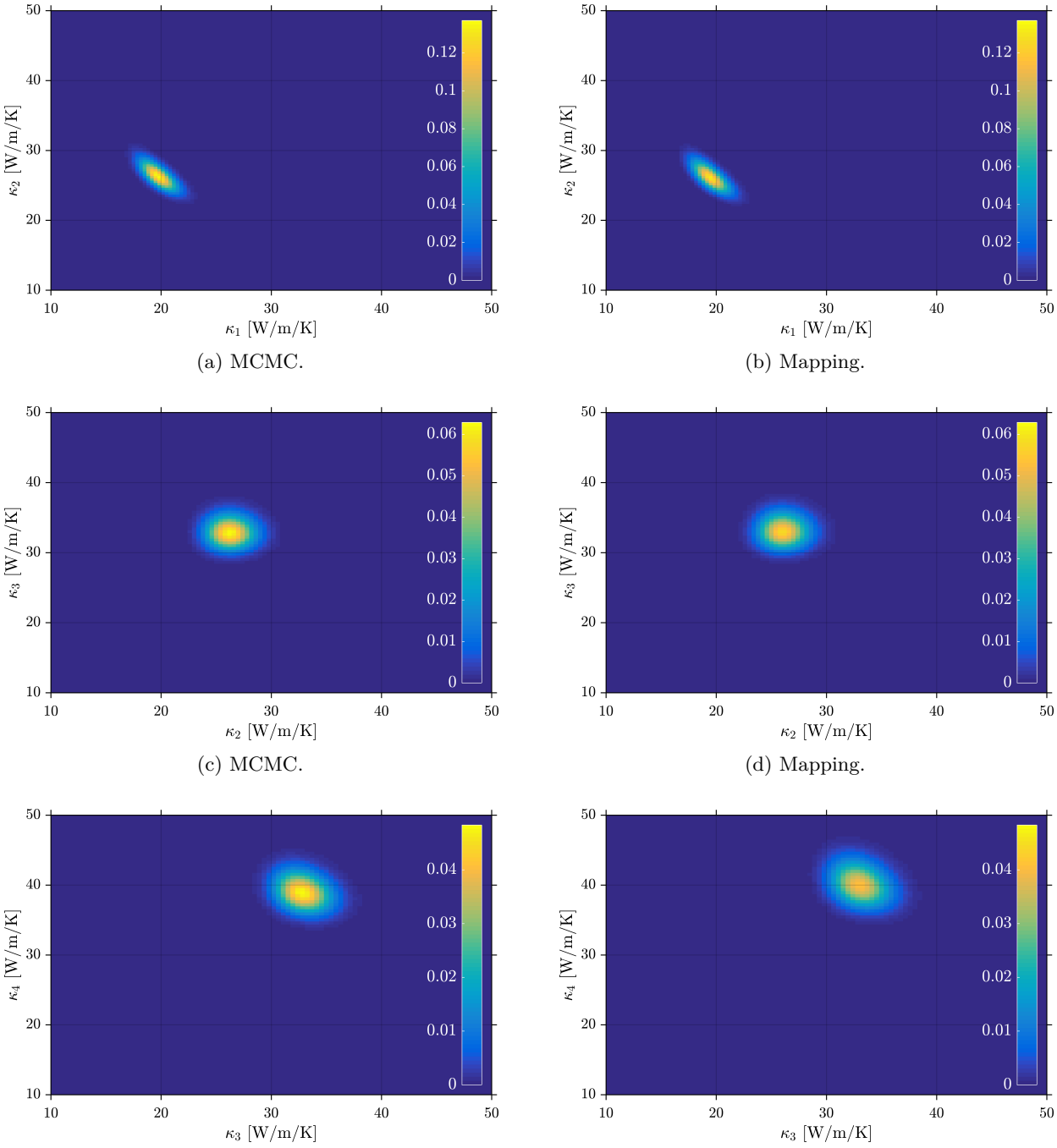
Figure 9.4: One-dimensional posterior marginals.

We also investigate some of the two-dimensional posterior marginals. A collection of bivariate histograms of those marginals can be found in Fig. 9.5. The results obtained from MCMC sampling are located on the left side, the ones from the prior transformation are shown on the right. In Figs. 9.5(a) and 9.5(b) the marginal $\pi(\kappa_1, \kappa_2|\boldsymbol{y})$ is visualized. Similarly, Figs. 9.5(c) and 9.5(d) contain $\pi(\kappa_2, \kappa_3|\boldsymbol{y})$ while Figs. 9.5(e) and 9.5(f) show $\pi(\kappa_3, \kappa_4|\boldsymbol{y})$. Transformation-based inference leads to two-dimensional posterior marginals that seem to be slightly flattened out as compared to their MCMC solutions.

The consideration of the sign of the Jacobian determinant in the course of the BFGS iterations $\iota = 0, \dots, I$ over the prior range provides some interesting insights. We find that the Jacobian $\det J_{T^{\triangle}_{\hat{\boldsymbol{a}}}}(\boldsymbol{x}^{(l)}) < 0$ of the finally computed map is negative for all prior samples $\boldsymbol{x}^{(l)}$ with $l = 1, \dots, L$. This indicates that the computed map $T^{\triangle}_{\hat{\boldsymbol{a}}}$ is indeed invertible over large proportions of the input space that are covered well by the prior. One might think that this justifies Eqs. (9.2) and (9.3), which only hold for invertible maps, in retrospect. However, while the algorithm is initialized such that the Jacobian $\det J_{T^{\triangle}_{\boldsymbol{a}_0}} > 0$ is positive in the beginning, it takes on positive and negative values $\det J_{T^{\triangle}_{\boldsymbol{a}_\iota}} \gtrless 0$ after some intermediary iterations. For such intermediate maps, the Jacobian determinant has zeros in regions of the parameter space that accumulate most prior mass. This signifies that these maps cannot be inverted globally.

The transformation and the MCMC results can be also compared by reference to the first posterior moments. In Table 9.4 the conditional means, standard deviations and linear correlations are summarized in tabular form. Apart from the correlation coefficients, all results are given in units of $[\kappa] = \mathrm{W/m/K}$. While the MCMC reference values are statistical sample approximations, transformation-based inference allows us to calculate the moments from the coefficients of the transport map through Eqs. (9.19) and (9.20). It is seen that the transportation manages to characterize the posterior in terms of its first statistical moments. The expected values $\mathbb{E}[\kappa_i|\boldsymbol{y}]$ are reproduced satisfactorily for $i = 1, \dots, 4$. As measured by the standard deviations $\mathrm{Std}[\kappa_i|\boldsymbol{y}] = \mathrm{Var}[\kappa_i|\boldsymbol{y}]^{1/2}$, the transformed distribution tends to overestimate the spread of the posterior slightly. The correlations $\rho[\kappa_i, \kappa_j|\boldsymbol{y}] = \mathrm{Cov}[\kappa_i, \kappa_j|\boldsymbol{y}]/\mathrm{Std}[\kappa_i|\boldsymbol{y}]/\mathrm{Std}[\kappa_j|\boldsymbol{y}]$ for $i, j = 1, \dots, 4$ are captured well. We conclude that, all in all, the prior has been successfully transformed into the posterior by the low-order transport map.

Lastly we investigate the model evidence Z. It can be estimated by brute-force MC simulation on the one hand. On the other hand, one may use the relation $Z \geq \exp(\mathcal{G}(T))$ that emerged in the context of Eqs. (9.11) and (9.12) in order to approximate the model evidence. After maximizing the sample average function in Eq. (9.21) as described by Eq. (9.22), $Z \approx \exp(\hat{\mathcal{G}}(\hat{\boldsymbol{a}}))$ may serve as a biased estimator of the model evidence. The convergence of this quantity in the optimization was already monitored in Table 9.2. With the abovementioned procedures, the model evidence is determined as $Z_{\mathrm{MC}} = 1.97 \times 10^{-5}$ and $Z_{\mathrm{Map}} = \exp(\hat{\mathcal{G}}(\hat{\boldsymbol{a}})) = 9.86 \times 10^{-5}$. Even though $\exp(\mathcal{G}(T))$ establishes a lower bound of the evidence in theory, $\exp(\hat{\mathcal{G}}(\hat{\boldsymbol{a}}))$ practically overestimates the MC reference solution. Contrary to the estimation of the first posterior moments that works rather well, the computation of the model evidence in inferential transportation seems to be more problematic.



(a) MCMC. (b) Mapping.

(c) MCMC. (d) Mapping.

(e) MCMC. (f) Mapping.

Figure 9.5: Two-dimensional posterior marginals.

Table 9.4: Posterior summaries.

	$\mathbb{E}[\kappa_1\|\boldsymbol{y}]$	$\mathbb{E}[\kappa_2\|\boldsymbol{y}]$	$\mathbb{E}[\kappa_3\|\boldsymbol{y}]$	$\mathbb{E}[\kappa_4\|\boldsymbol{y}]$	$\mathrm{Std}[\kappa_1\|\boldsymbol{y}]$	$\mathrm{Std}[\kappa_2\|\boldsymbol{y}]$	$\mathrm{Std}[\kappa_3\|\boldsymbol{y}]$
MCMC	19.78	26.36	32.94	39.06	1.11	1.50	1.69
Mapping	19.52	26.18	33.16	40.18	1.12	1.55	1.78
	$\mathrm{Std}[\kappa_4\|\boldsymbol{y}]$	$\rho[\kappa_1,\kappa_2\|\boldsymbol{y}]$	$\rho[\kappa_1,\kappa_3\|\boldsymbol{y}]$	$\rho[\kappa_1,\kappa_4\|\boldsymbol{y}]$	$\rho[\kappa_2,\kappa_3\|\boldsymbol{y}]$	$\rho[\kappa_2,\kappa_4\|\boldsymbol{y}]$	$\rho[\kappa_3,\kappa_4\|\boldsymbol{y}]$
MCMC	2.01	−0.38	−0.22	−0.02	−0.03	−0.72	−0.36
Mapping	2.25	−0.39	−0.24	−0.01	−0.01	−0.72	−0.38

9.6 Summary and conclusion

An approach to Bayesian inference based on transport maps was investigated in this chapter. The prior and the posterior were coupled based on an appropriate change of variables. Practically this was done by minimizing the Kullback–Leibler divergence of the back-transformed posterior from the prior. The optimization problem faced was regularized in the framework of optimal transportation theory. A triangular map with polynomial components was used to parameterize the sought transformation. The upside of the technique is that it works in principle and indeed establishes a doable option for probabilistic inference. Due to the lack of fundamental alternatives to conventional Markov chain Monte Carlo techniques, this is a very strong point that makes further research attractive and needful.

On the downside, finding an appropriate transformation comes at a high computational price. The optimization problem posed involves expectations under the prior distribution over the likelihood function. Even though it was found that low-degree polynomials suffice in order to transform between the prior and the posterior distribution, a high number of samples from the prior are required for approximating the corresponding utility function. This imposes an immense number of likelihood evaluations that are necessary for each call to the utility in every algorithm iteration.

A host of open questions has been given rise to. An in-depth understanding of the optimization problem and its discretization would support the choice of well-suited optimizers and their algorithmic parameters. It would be helpful to have a solid criterion assisting in setting the sample size of the Monte Carlo approximation and the polynomial degree of the triangular map. While the model evidence establishes an upper bound of the utility function, it cannot serve as a target value for assessing convergence, since we do not actually know it. This is exacerbated because we obtain an acceptable solution to an approximate problem at most by maximizing the sample average function. Vice versa, the eventually obtained maximum of the utility function does not necessarily serve well as an estimator of the log–model evidence. It is biased downwards. As for the Monte Carlo sample size used in the sample average approximation and also the maximal polynomial degree, we had to rely on heuristic criteria and checks against the results from a Markov chain Monte Carlo procedure.

Another question that was brought up is how one can enforce invertibility of the transformation. This is necessary in order to warrant the correctness of the change of variables formula that the optimization objective builds on. The issue was recklessly but wittingly ignored in the current approach. Moreover, it would be desirable to restrict the search for transformations to such ones that comply with possibly existing prior constraints and do not map out of the permissible range.

Beyond detail improvements of the transport map approach, one could envisage the combination with sampling-based approaches. Markov chain Monte Carlo sampling could be accelerated by transforming a standard proposal into a non-standard distribution that strongly overlaps with the posterior. Looking at it the other way around, one could also transform a complex posterior into a simpler target distribution. Indeed there is ongoing research in these directions [32, 33].

References

[1] T. A. El Moselhy and Y. M. Marzouk. “Bayesian inference with optimal maps”. In: *Journal of Computational Physics* 231.23 (2012), pp. 7815–7850. DOI: 10.1016/j.jcp.2012.07.022.

[2] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. “Sampling via Measure Transport: An Introduction”. In: *Handbook of Uncertainty Quantification*. Ed. by R. Ghanem, D. Higdon, and H. Owhadi. Cham, Switzerland: Springer International Publishing, 2016. DOI: 10.1007/978-3-319-11259-6_23-1.

[3] S. Reich and C. J. Cotter. “Ensemble filter techniques for intermittent data assimilation”. In: *Large Scale Inverse Problems: Computational Methods and Applications in the Earth Sciences*. Ed. by M. Cullen, M. A. Freitag, S. Kindermann, and R. Scheichl. Radon Series on Computational and Applied Mathematics 13. Berlin, Germany: Walter de Gruyter Verlag, 2013, pp. 91–134.

[4] P. J. Van Leeuwen, Y. Cheng, and S. Reich. *Nonlinear Data Assimilation*. Frontiers in Applied Dynamical Systems: Reviews and Tutorials 2. Cham, Switzerland: Springer International Publishing, 2015. DOI: 10.1007/978-3-319-18347-3.

[5] J. B. Nagel and B. Sudret. “Optimal Transportation for Bayesian Inference in Engineering”. In: *International Symposium on Reliability of Engineering Systems (SRES 2015)*. Hangzhou, China, October 2015.

[6] H. Thorisson. *Coupling, Stationarity, and Regeneration*. Probability and its Applications. New York: Springer-Verlag, 2000.

[7] T. Lindvall. *Lectures on the Coupling Method*. Dover Books on Mathematics. Mineola, New York, USA: Dover Publications, Inc., 2002.

[8] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics 58. Providence, Rhode Island, USA: American Mathematical Society (AMS), 2003. DOI: 10.1090/gsm/058.

[9] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften 338. Springer-Verlag Berlin Heidelberg, 2008. DOI: 10.1007/978-3-540-71050-9.

[10] Y. Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics* 44.4 (1991), pp. 375–417. DOI: 10.1002/cpa.3160440402.

[11] R. J. McCann. "Existence and uniqueness of monotone measure-preserving maps". In: *Duke Mathematical Journal* 80.2 (1995), pp. 309–323. DOI: 10.1215/S0012-7094-95-08013-2.

[12] G. Carlier, A. Galichon, and F. Santambrogio. "From Knothe's Transport to Brenier's Map and a Continuation Method for Optimal Transport". In: *SIAM Journal on Mathematical Analysis* 41.6 (2010), pp. 2554–2576. DOI: 10.1137/080740647.

[13] N. Bonnotte. "From Knothe's Rearrangement to Brenier's Optimal Transport Map". In: *SIAM Journal on Mathematical Analysis* 45.1 (2013), pp. 64–87. DOI: 10.1137/120874850.

[14] E. S. Soofi, H. Zhao, and D. L. Nazareth. "Information measures". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.1 (2010), pp. 75–86. DOI: 10.1002/wics.62.

[15] N. Ebrahimi, E. S. Soofi, and R. Soyer. "Information Measures in Perspective". In: *International Statistical Review* 78.3 (2010), pp. 383–412. DOI: 10.1111/j.1751-5823.2010.00105.x.

[16] A. Prékopa. *Stochastic Programming*. Mathematics and Its Applications 324. Dordrecht, Netherlands: Kluwer Academic Publishers, 1995. DOI: 10.1007/978-94-017-3087-7.

[17] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. 2nd ed. MOS-SIAM Series on Optimization. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 2014.

[18] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello. "The Sample Average Approximation Method for Stochastic Discrete Optimization". In: *SIAM Journal on Optimization* 12.2 (2002), pp. 479–502. DOI: 10.1137/S1052623499363220.

[19] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609. DOI: 10.1137/070704277.

[20] J. Wang and N. Zabaras. "A Bayesian inference approach to the inverse heat conduction problem". In: *International Journal of Heat and Mass Transfer* 47.17–18 (2004), pp. 3927–3941. DOI: 10.1016/j.ijheatmasstransfer.2004.02.028.

[21] J. P. Kaipio and C. Fox. "The Bayesian Framework for Inverse Problems in Heat Transfer". In: *Heat Transfer Engineering* 32.9 (2011), pp. 718–753. DOI: 10.1080/01457632.2011.525137.

[22] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. 3rd ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2006. DOI: 10.1002/0471787779.

[23] W. Sun and Y.-X. Yuan. *Optimization Theory and Methods: Nonlinear Programming*. Springer Optimization and Its Applications 1. New York: Springer, 2006. DOI: 10.1007/b106451.

[24] P. Zörnig. *Nonlinear Programming: An Introduction*. De Gruyter Textbook. Berlin, Germany: Walter de Gruyter, 2014.

[25] C. G. Broyden. "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations". In: *Journal of the Institute of Mathematics and its Applications* 6.1 (1970), pp. 76–90. DOI: 10.1093/imamat/6.1.76.

[26] C. G. Broyden. "The Convergence of a Class of Double-rank Minimization Algorithms 2. The New Algorithm". In: *Journal of the Institute of Mathematics and its Applications* 6.3 (1970), pp. 222–231. DOI: 10.1093/imamat/6.3.222.

[27] R. Fletcher. "A new approach to variable metric algorithms". In: *The Computer Journal* 13.3 (1970), pp. 317–322. DOI: 10.1093/comjnl/13.3.317.

[28] D. Goldfarb. "A Family of Variable-Metric Methods Derived by Variational Means". In: *Mathematics of Computation* 24.109 (1970), pp. 23–26. DOI: 10.1090/S0025-5718-1970-0258249-6.

[29] D. F. Shanno. "Conditioning of Quasi-Newton Methods for Function Minimization". In: *Mathematics of Computation* 24.111 (1970), pp. 647–656. DOI: 10.1090/S0025-5718-1970-0274029-X.

[30] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. New York: Springer, 2006. DOI: 10.1007/978-0-387-40065-5.

[31] A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. 2nd ed. Natural Computing Series. Springer-Verlag Berlin Heidelberg, 2015. DOI: 10.1007/978-3-662-44874-8.

[32] M. Parno and Y. Marzouk. *Transport map accelerated Markov chain Monte Carlo*. 2014. arXiv: 1412.5492 [stat.CO].

[33] M. D. Parno. "Transport maps for accelerated Bayesian computation". PhD thesis. Cambridge, Massachusetts, USA: Massachusetts Institute of Technology (MIT), 2015.

Chapter 10

Hydrological black-box model calibration

This chapter deals with the Bayesian calibration of a dynamical urban drainage simulator. The process-based simulation of water systems is typically expensive and yet highly uncertain [1]. That is why hydrological model calibration is an extremely important and difficult task. We use a combination of advanced methods in order to estimate the unknown model parameters and to quantify the measurement and prediction errors. Principal component analysis is used for purposes of dimension reduction of the model outputs that constitute a whole times series. The accordingly reduced outputs are then metamodeled as a function of the unknowns based on sparse polynomial chaos expansions. Eventually the posterior distribution of the unknown parameters of the hydrological and the error model is sampled via Markov chain Monte Carlo.

The process-oriented hydrological simulator predicts the outflow from a catchment area that receives rainfall. It has been developed at the Swiss Federal Institute of Aquatic Science and Technology (Eawag), where it was used in the PhD dissertation of D. Machac [2]. By courtesy of Eawag we have access to the results of roughly two thousand training runs of the simulator for a single rainfall event. Moreover, about six hundred measurements of the time-varying runoff at a single outlet during the event were made available. This describes a realistic black-box situation where the abovementioned hybridization of techniques for compression, metamodeling and calibration permits a synthesis of the supplied information.

Measurement uncertainties and modeling errors are explicitly considered in two different Bayesian models. Parametric uncertainties in the hydrological inputs are the main focus of both models. They differ, however, in the degree of sophistication of how the inevitable deviation of the model predictions from the measurement data is represented. The first simple model only acknowledges independent measurement noise, while the second model also accounts for random error correlation and systematic model discrepancy.

The present chapter is organized as follows. A more detailed overview of the problem setup and the available information is provided in Section 10.1. The construction of the hydrological emulator is described in Section 10.2. Bayesian parameter estimation and predictive model correction are performed in Section 10.3. Finally it is summarized and concluded in Section 10.4.

10.1 Problem setup

As elsewhere, in hydrology one distinguishes between physical process-based and purely data-driven modeling approaches [3–5]. Even if a model is primarily based on physical principles, the parameters and predictions have to be respectively calibrated and corrected in conjunction with experimental data. This is the goal of this chapter. The physics-oriented hydrological model under consideration predicts the outflow from an urban drainage basin in a precipitation event. Dynamical simulations of the runoff are based on a series of rainfall measurements. Some further model inputs are unknown and shall be calibrated with time series data of the outflow. Bounds of these unknowns are established and their prior distributions are prescribed. If the model could be run for arbitrarily chosen values of the inputs, the uncertain parameters could be identified with the methods previously discussed in this thesis. However, this is not the case, because the model is not available to us in an executable form.

We only have the results of a limited number of simulator runs that were performed for certain rainfall data and uniformly varying values of the unknown parameters. In this situation we have to construct an operational emulator first. The originally performed model runs are fed into a metamodeling procedure for that purpose.

After a surrogate model is obtained, one can use it in order to "interpolate" between the design points in continuous fashion. This way the simulator output can be at least computed approximately for arbitrary values of the unknowns. Subsequently Bayesian inference proceeds as usual.

We start with a brief description of the urban drainage model and the measurement data. More details can be taken from the fourth chapter of [2]. The storm water management model (SWMM) is a dynamic rainfall-runoff simulation program for urban areas [6]. It can be used to predict the runoff from a catchment area during and shortly after a rainfall event. A model of the drainage basin of Adliswil, a municipality in the canton of Zürich in the northeast of Switzerland, was created with the SWMM. In Fig. 10.1 a map is provided that shows the surrounding area of the size 5 km × 3 km. The SWMM implementation models about 160 ha of this area, i.e. approximately ten percent. Roughly speaking, this SWMM implementation uses one hundred sub-catchments that are linked through five hundred channels.

Figure 10.1: Adliswil, Switzerland (1:50000). From OpenStreetMap under CC BY-SA 2.0. © OpenStreetMap contributors.

All sub-catchments and interconnections have their own unknown parameters. This amounts to a fairly large number of unknowns that is reduced by considering spatial averages only, i.e. physical quantities of the same type are averaged over the sub-catchments or channels. Moreover, the parameter classes are normalized so as to be dimensionless and to lie in between reasonable bounds. A compilation of the obtained scaled parameters $x_i \in \mathcal{D}_{x_i}$ for $i = 1, \ldots, 8$ and their bounded domains $\mathcal{D}_{x_i} = [\underline{x}_i, \overline{x}_i]$ is presented in Table 10.1 below. The physical quantities described and their unscaled spatial averages are also provided. While the first seven parameters relate to the sub-catchments, only the last one characterizes the pipes.

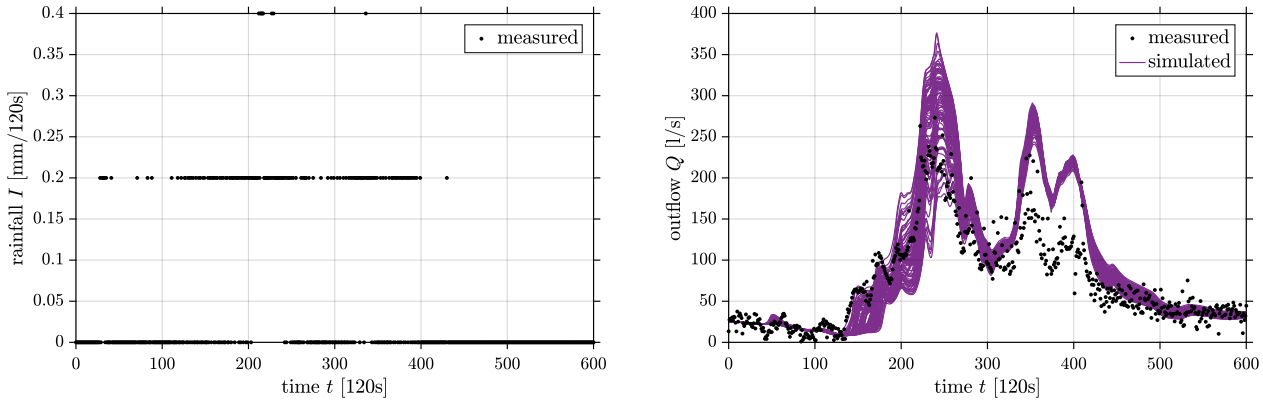
A single 15-hour rainfall event is considered that had occurred on May 28, 2013. Time is denoted as t in the following. The experiment extends over a period with $t/120\,\mathrm{s} \in [0, 600]$. Measurements of the varying rainfall intensity I and the catchment outflow Q are taken in regular intervals of two minutes over the full duration. For $i = 0, \ldots, 600$ the time instances of the observations are denoted as t_i. Both rainfall and outflow measurements were made at single locations within the drainage basin, e.g. the outflow was measured at the wastewater treatment plant. In Fig. 10.2 the available data are summarized. The observations of the rainfall intensity $I(t_i)$ are indicated by the black dots in Fig. 10.2(a). Similarly, the recorded outflows $Q(t_i)$ at the sewage treatment plant are shown in Fig. 10.2(b).

Beyond the observational data just described, the results of approximately two thousand runs of the SWMM simulator are available. These will constitute the training runs for the computation of the surrogate model in the next section. They were conducted for the given rainfall data shown in Fig. 10.2(a) and uniformly distributed values of the uncertain hydrological parameters. A hundred trajectories from these computer simulations are depicted in Fig. 10.2(b). They can be compared to the actually measured runoffs in the same plot.

Table 10.1: Hydrological model parameters.

x_i	$\mathcal{D}_{x_i}$	Physical parameter	Spatial average
x_1	$[0.5, 1.1]$	Percentage of the impervious area	$36\,\%$
x_2	$[0.5, 1.5]$	Characteristic width of the overland flow path	$35.7\,\mathrm{m}$
x_3	$[0.5, 1.5]$	Slope of the sub-catchments	$11.4\,\%$
x_4	$[0.5, 1.5]$	Depression storage height of the impervious area	$2\,\mathrm{mm}$
x_5	$[0.5, 1.5]$	Manning roughness coefficient of the impervious area	$0.12\,\mathrm{s}\cdot\mathrm{m}^{-1/3}$
x_6	$[0.5, 1.5]$	Depression storage height of the pervious area	$2\,\mathrm{mm}$
x_7	$[0.5, 1.5]$	Percentage of the impervious area without depression storage	$19.04\,\%$
x_8	$[1.0, 1.5]$	Manning roughness coefficient of the channels	$0.012\,\mathrm{s}\cdot\mathrm{m}^{-1/3}$

The model manages to capture the main trends and characteristics of the data. In the time interval $t/120\,\mathrm{s} \in [150, 200]$ it seems to slightly underpredict the outflow, though. An even stronger systematic discrepancy is detected for the time span $t/120\,\mathrm{s} \in [250, 500]$ during which the outflow is overpredicted. It is also noticed that the model predictions for different values of the uncertain inputs do not differ significantly, i.e. they cannot be discriminated very well by their ability to trace the data. This is especially obvious in the second half of the experiment with $t/120\,\mathrm{s} \in [300, 600]$. Here, the mismatch between the data and the model predictions is apparently dominated by systematic errors and random noise, rather than by variations of the model inputs.



(a) Rainfall intensity data.

(b) Observed and simulated outflow.

Figure 10.2: Experimental data.

10.2 Metamodeling

Dynamic simulators can be emulated in a purely statistics-based manner, e.g. by conditioning a Gaussian process prior on the experimental design. Mechanism-based approaches try to enhance the emulation through an appropriate incorporation of the available physical understanding [7–9]. In particular, the solution of a simplified problem is incorporated into the prior and subsequently corrected so as to emulate the full simulator. An application of this approach to the Adliswil watershed can be found in [10].

We pursue an alternative strategy. First, the model output dimensionality is reduced through principal component analysis. Then, sparse polynomial chaos expansions are computed for the main components as functions of the uncertain inputs. Last, the obtained expansions are combined in order to obtain a surrogate model for the full time series of the outflow.

10.2.1 Computational model

The SWMM implementation of the catchment predicts a complete time series of the outflow at the wastewater treatment plant throughout the precipitation event. Provided that one inputs the rainfall intensity over the duration of the experiment, the model only acts as a function of the uncertain input parameters listed in Table 10.1. We gather the unknown model input parameters in a vector $\boldsymbol{x} = (x_1, \ldots, x_M)^\top$ with $M = 8$.

Similarly we proceed for the rainfall data $\boldsymbol{d} = (d_0, \ldots, d_N)^\top$ with $N = 600$. For $i = 0, \ldots, N$ we have introduced $d_i = I(t_i)$ for the observed rainfall intensities at the measurement time instances t_i.

The numerical model predicts the vector $\tilde{\boldsymbol{y}} = (\tilde{y}_0, \ldots, \tilde{y}_N)^\top$ whose entries are the outflows $\tilde{y}_i = \tilde{Q}(t_i)$ at the times t_i. All in all, $\tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{x}, \boldsymbol{d})$ reflects the structure of the hydrological simulations. Since we only consider a single precipitation event and disregard errors in the rainfall data, we absorb the dependence on the rainfall into the definition of the forward model $\mathcal{M}_{\boldsymbol{d}}$ by

$$\tilde{\boldsymbol{y}} = \mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x}). \tag{10.1}$$

We now switch to a probabilistic formulation, where the inputs are assumed to be independent and $[\underline{x}_i, \overline{x}_i]$-valued random variables X_i with uniform distributions $X_i \sim \mathcal{U}(x_i | \underline{x}_i, \overline{x}_i)$ for $i = 1, \ldots, M$. The random vector $\boldsymbol{X} = (X_1, \ldots, X_M)^\top$ is then distributed according to

$$\boldsymbol{X} \sim \prod_{i=1}^{M} \mathcal{U}(x_i | \underline{x}_i, \overline{x}_i). \tag{10.2}$$

This distribution represents some kind of input uncertainty. When the model $\mathcal{M}_{\boldsymbol{d}}$ is applied to the random inputs $\boldsymbol{X}$, the output uncertainty is described by the $\mathbb{R}^{N+1}$-valued random response vector

$$\tilde{\boldsymbol{Y}} = \mathcal{M}_{\boldsymbol{d}}(\boldsymbol{X}). \tag{10.3}$$

As already mentioned before, the original implementation of this simulator is only available to us through $K = 2048$ training runs in total. We cannot execute it for arbitrary values of the inputs. Yet we have access to realizations $\boldsymbol{x}^{(k)}$ of the input variables in Eq. (10.2) for $k = 1, \ldots, K$ and the corresponding realizations $\tilde{\boldsymbol{y}}^{(k)} = \mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x}^{(k)})$ of Eq. (10.3). The inputs were obtained by Latin hypercube sampling [11] in two chunks of 1024 samples each. Altogether they constitute the experimental design $\mathcal{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)})$. Moreover, the responses are collected into the data matrix

$$\mathcal{Y} = \begin{pmatrix} \tilde{\boldsymbol{y}}^{(1)\top} \\ \tilde{\boldsymbol{y}}^{(2)\top} \\ \vdots \\ \tilde{\boldsymbol{y}}^{(K)\top} \end{pmatrix} = \begin{pmatrix} \tilde{y}_0^{(1)} & \tilde{y}_1^{(1)} & \cdots & \tilde{y}_N^{(1)} \\ \tilde{y}_0^{(2)} & \tilde{y}_1^{(2)} & \cdots & \tilde{y}_N^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{y}_0^{(K)} & \tilde{y}_1^{(K)} & \cdots & \tilde{y}_N^{(K)} \end{pmatrix}. \tag{10.4}$$

With that said, we have to embrace a black-box perspective on the present problem. The information contained in the training runs is used in the computation of a surrogate simulator, i.e. a metamodeling problem with $M = 8$ inputs, $N + 1 = 601$ outputs and $K = 2048$ training runs is posed. The problem is solved with a multivariate extension of the polynomial chaos–based methods described in Sections 2.3 and 2.4.

10.2.2 Principal components

Note that a preliminary discussion of the multivariate case was already provided in Section 2.5. The coordinates of the model output with respect to a certain reference system, e.g. the canonical basis, can be considered individually. In our case, this would require to handle about six hundred different metamodels at the same time. That is inconvenient and involves a high degree of redundancy, i.e. the simulation outputs at contiguous times are highly correlated.

To find a remedy one can choose a basis that is qualified for purposes of dimension reduction and data compression. Here we use *principal component analysis* (PCA) [12–14] to that end. While this technique is mainly used for compressing big real-word data sets with many features, it can be similarly used for reducing the model output in the context of computer simulations [15, 16]. We start by discussing the population PCA for a random vector, which is just the discrete variant of the *Karhunen–Loève* (KL) *expansion* of a stochastic process [17]. Afterwards the empirical sample PCA is recalled.

Consider the random vector $\tilde{\boldsymbol{Y}}$ with mean $\boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}} = \mathbb{E}[\tilde{\boldsymbol{Y}}]$ and covariance matrix $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}} = \mathrm{Cov}[\tilde{\boldsymbol{Y}}] = \mathbb{E}[(\tilde{\boldsymbol{Y}} - \boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}})(\tilde{\boldsymbol{Y}} - \boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}})^\top]$. Since $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}}$ is symmetric and positive definite, one can find linearly independent eigenvectors $\boldsymbol{\phi}_i$ with positive eigenvalues $\lambda_i > 0$ for $i = 0, \ldots, N$. The characteristic vectors and values satisfy

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}} \boldsymbol{\phi}_i = \lambda_i \boldsymbol{\phi}_i. \tag{10.5}$$

Eigenvectors corresponding to distinct eigenvalues are orthogonal anyway, while they can be always chosen as such for degenerate eigenvalues. We assume that the eigenvalues are arranged in decreasing order $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_N$ and that eigenvectors are normalized such that $\boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j = \delta_{ij}$ for $i, j = 0, \ldots, N$. Leaving degeneracy aside, this way the eigenvectors are uniquely defined up to a multiplication by -1.

The set of eigenvectors constitutes an orthonormal basis of $\mathbb{R}^{N+1} = \operatorname{span}(\{\boldsymbol{\phi}_i\}_{i=0}^N)$. One can define the orthogonal matrix $\boldsymbol{\Phi} = (\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N)$ with $\boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \boldsymbol{I}$. It diagonalizes the covariance matrix by

$$\boldsymbol{\Phi}^\top \boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}} \boldsymbol{\Phi} = \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_0 & 0 & \ldots & 0 \\ 0 & \lambda_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_N \end{pmatrix}. \tag{10.6}$$

Vice versa, one obtains the spectral eigendecomposition of the covariance matrix $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^\top = \sum_{i=0}^N \lambda_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$. Now consider the orthogonal transformation

$$\tilde{\boldsymbol{Z}} = \boldsymbol{\Phi}^\top (\tilde{\boldsymbol{Y}} - \boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}}). \tag{10.7}$$

The linearly transformed random vector has mean zero $\mathbb{E}[\tilde{\boldsymbol{Z}}] = \boldsymbol{0}$ and the diagonal covariance matrix $\operatorname{Cov}[\tilde{\boldsymbol{Z}}] = \mathbb{E}[\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top] = \boldsymbol{\Lambda}$, i.e. it has been centered and decorrelated. Independence is not necessarily implied thereby, though, the special case involving Gaussianity forms an exception. The back-transformation reads

$$\tilde{\boldsymbol{Y}} = \boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}} + \boldsymbol{\Phi}\tilde{\boldsymbol{Z}} = \boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}} + \sum_{i=0}^N \tilde{Z}_i \boldsymbol{\phi}_i. \tag{10.8}$$

This is the discrete KL expansion of the random vector $\tilde{\boldsymbol{Y}}$. The random variables $\tilde{Z}_i = \boldsymbol{\phi}_i^\top (\tilde{\boldsymbol{Y}} - \boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}})$ for $i = 0, \ldots, N$ are called the *principal components*.

Define the *total variance* of $\tilde{\boldsymbol{Y}}$ as the sum $\sum_{i=0}^N \operatorname{Var}[\tilde{Y}_i]$ of the individual variances of $\tilde{Y}_i$. The orthogonal transformation preserves the total variance in the sense that

$$\sum_{i=0}^N \operatorname{Var}[\tilde{Y}_i] = \operatorname{tr}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}}) = \operatorname{tr}(\boldsymbol{\Lambda}) = \sum_{i=0}^N \operatorname{Var}[\tilde{Z}_i] = \sum_{i=0}^N \lambda_i. \tag{10.9}$$

This follows from the invariance of the trace under cyclic permutations. The KL expansion is optimal with respect to compaction of the total variance. Consider keeping only the first $N' + 1 \leq N + 1$ terms in

$$\tilde{\boldsymbol{Y}} \approx \boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}} + \sum_{i=0}^{N'} \tilde{Z}_i \boldsymbol{\phi}_i. \tag{10.10}$$

This is the expansion that contains most of the total variance with $N' + 1$ terms. The number of terms is often chosen such that at least a predetermined fraction $\sum_{i=0}^{N'} \lambda_i / \sum_{i=0}^N \lambda_i$ of the total variance is explained.

The sample PCA functions in exactly the same way for independent realizations as the population PCA does for random vectors. Instead of the exact mean $\boldsymbol{\mu}_{\tilde{\boldsymbol{Y}}} = \mathbb{E}[\tilde{\boldsymbol{Y}}]$ and covariance matrix $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Y}}} = \operatorname{Cov}[\tilde{\boldsymbol{Y}}]$, one considers their empirical estimates for the sample $\mathcal{Y} = (\tilde{\boldsymbol{y}}^{(1)}, \ldots, \tilde{\boldsymbol{y}}^{(K)})^\top$ of realizations from $\tilde{\boldsymbol{Y}}$. They are given as

$$\overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}} = \frac{1}{K} \sum_{k=1}^K \tilde{\boldsymbol{y}}^{(k)}, \quad \overline{\boldsymbol{\Sigma}}_{\tilde{\boldsymbol{Y}}} = \frac{1}{K-1} \sum_{k=1}^K (\tilde{\boldsymbol{y}}^{(k)} - \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}})(\tilde{\boldsymbol{y}}^{(k)} - \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}})^\top. \tag{10.11}$$

For $i = 0, \ldots, N$ the eigenvectors $\overline{\boldsymbol{\phi}}_i$ and eigenvalues $\overline{\lambda}_i$ of the empirical covariance fulfill $\overline{\boldsymbol{\Sigma}}_{\tilde{\boldsymbol{Y}}} \overline{\boldsymbol{\phi}}_i = \overline{\lambda}_i \overline{\boldsymbol{\phi}}_i$. The eigenvalues are arranged in the descending order $\overline{\lambda}_0 \geq \overline{\lambda}_1 \geq \ldots \geq \overline{\lambda}_N$.

Then one finds the smallest $N' \leq N$ for which the proportion $\sum_{i=0}^{N'} \overline{\lambda}_i / \sum_{i=0}^N \overline{\lambda}_i$ of the total empirical variance is larger or at least equal than a prespecified threshold. The matrix $\overline{\boldsymbol{\Phi}}_{N'} = (\overline{\boldsymbol{\phi}}_0, \overline{\boldsymbol{\phi}}_1, \ldots, \overline{\boldsymbol{\phi}}_{N'})$ is composed and for $k = 1, \ldots, K$ one defines

$$\tilde{\boldsymbol{z}}^{(k)} = \overline{\boldsymbol{\Phi}}_{N'}^\top (\tilde{\boldsymbol{y}}^{(k)} - \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}}). \tag{10.12}$$

This is the reduced PCA representation of $\tilde{\boldsymbol{y}}^{(k)}$ in terms of the empirical principal components $\tilde{z}_i^{(k)} = \overline{\boldsymbol{\phi}}_i^\top (\tilde{\boldsymbol{y}}^{(k)} - \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}})$ for $i = 0, \ldots, N'$. The data set is compressed while retaining most of the total variation by

$$\mathcal{Z} = \begin{pmatrix} \tilde{\boldsymbol{z}}^{(1)\top} \\ \tilde{\boldsymbol{z}}^{(2)\top} \\ \vdots \\ \tilde{\boldsymbol{z}}^{(K)\top} \end{pmatrix} = \begin{pmatrix} \tilde{z}_0^{(1)} & \tilde{z}_1^{(1)} & \ldots & \tilde{z}_{N'}^{(1)} \\ \tilde{z}_0^{(2)} & \tilde{z}_1^{(2)} & \ldots & \tilde{z}_{N'}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{z}_0^{(K)} & \tilde{z}_1^{(K)} & \ldots & \tilde{z}_{N'}^{(K)} \end{pmatrix}. \tag{10.13}$$

The compression of the data is still lossy, but one can reconstruct the originally observed samples for $k = 1, \ldots, K$ approximately as

$$\tilde{\boldsymbol{y}}^{(k)} \approx \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}} + \overline{\boldsymbol{\Phi}}_{N'} \tilde{\boldsymbol{z}}^{(k)} = \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}} + \sum_{i=0}^{N'} \tilde{z}_i^{(k)} \overline{\boldsymbol{\phi}}_i. \tag{10.14}$$

10.2.3 Sparse expansion

Now we perform PCA to our sample of SWMM simulator responses $\mathcal{Y} = (\tilde{\boldsymbol{y}}^{(1)}, \dots, \tilde{\boldsymbol{y}}^{(K)})^\top$. Using $N' + 1 = 9$ principal components captures 99 % of the total variance of the signal. Hundred realizations contained in the compressed data set $\mathcal{Z} = (\tilde{\boldsymbol{z}}^{(1)}, \dots, \tilde{\boldsymbol{z}}^{(K)})^\top$ are visualized in the parallel coordinate plot in Fig. 10.3. These are the empirical principal components of the sample of training runs that were already shown in Fig. 10.2(b). It can be seen that the main components are centered around zero and ordered according to their individual contribution to the total variance.

With the underlying experimental design $\mathcal{X}$ and the reduced output data $\mathcal{Z}$, PCEs of the population principal components $\tilde{z}_i(\boldsymbol{X})$ are computed separately for $i = 0, \dots, N'$. Normalized multivariate Legendre polynomials $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{X})\}_{\boldsymbol{\alpha} \in \mathcal{A}_p}$ in the random inputs constitute the expansion basis for all PCEs. The total degree $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^{M} |\alpha_i| \leq p$ is limited to at most $p = 10$. Only terms with $\boldsymbol{\alpha} \in \mathcal{A}_p = \left\{\boldsymbol{\beta} \in \mathbb{N}^M : \|\boldsymbol{\beta}\|_1 \leq p\right\}$ are then kept in the PCEs

$$\tilde{z}_i(\boldsymbol{X}) \approx \sum_{\boldsymbol{\alpha} \in \mathcal{A}_p} \hat{a}_{i,\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}). \tag{10.15}$$

We use *least angle regression* (LAR) [18, 19] in order to compute the coefficients $\{\hat{a}_{i,\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A}_p}$ for each expansion with $i = 0, \dots, N'$. LAR is a powerful technique that promotes sparsity in the PCE coefficient vectors. Regressors are penalized such a way that only the most dominant ones are retained in the expansion. This allows us to mitigate the curse of dimensionality discussed in Section 2.6 to some degree. The algorithm has been proven very efficient in the context of metamodeling based on polynomial chaos [20]. We use our own inhouse implementation of LAR [21].

Cross validation is used to assess the generalization performance of the sparse PCEs. The normalized leave-one-out errors for expansions with $K = 1024$ and $K = 2048$ are reported in Table 10.2. As expected, the PCE with the richer experimental design generalizes better than the one with the poorer design for which the error is approximately twice as high. One can observe the general trend that the accuracy of the approximation decays with the order of the principal components. As it turns out, the hydrological model is indeed approximately sparse in the polynomial basis used. Only a small fraction of the total number of regressors is retained in each of the nine expansions.

Table 10.2: Normalized leave-one-out errors.

K	$\tilde{z}_0$	$\tilde{z}_1$	$\tilde{z}_2$	$\tilde{z}_3$	$\tilde{z}_4$	$\tilde{z}_5$	$\tilde{z}_6$	$\tilde{z}_7$	$\tilde{z}_8$
1024	2.58×10^{-5}	1.42×10^{-4}	3.94×10^{-4}	2.22×10^{-4}	3.37×10^{-3}	2.70×10^{-3}	7.35×10^{-3}	7.37×10^{-3}	1.11×10^{-2}
2048	1.49×10^{-5}	6.87×10^{-5}	1.73×10^{-4}	1.06×10^{-4}	1.41×10^{-3}	1.23×10^{-3}	2.56×10^{-3}	2.97×10^{-3}	3.16×10^{-3}

After the computation of a PCE for each principal component in $\tilde{\boldsymbol{z}}(\boldsymbol{X}) = (\tilde{z}_0(\boldsymbol{X}), \dots, \tilde{z}_{N'}(\boldsymbol{X}))^\top$, the random vector containing the model outputs is approximately represented as

$$\tilde{\boldsymbol{Y}} \approx \hat{\mathcal{M}}_p(\boldsymbol{X}) = \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}} + \sum_{i=0}^{N'} \tilde{z}_i(\boldsymbol{X}) \overline{\boldsymbol{\phi}}_i. \tag{10.16}$$

This expansion is henceforth used as a metamodel of the model output $\tilde{\boldsymbol{y}} \approx \hat{\mathcal{M}}_p(\boldsymbol{x}) = \overline{\boldsymbol{\mu}}_{\tilde{\boldsymbol{Y}}} + \sum_{i=0}^{N'} \tilde{z}_i(\boldsymbol{x}) \overline{\boldsymbol{\phi}}_i$ at arbitrary input values $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$. This is justified due to the mean-square convergence of the underlying PCEs. In Fig. 10.4 the simulated and emulated outflows are shown for three different input values in the experimental design. It is ascertained that the obtained metamodel is sufficiently accurate for parameter calibration purposes.

10.3 Bayesian calibration

We now turn towards model calibration. The goal is to infer the unknown hydrological parameters with the available outflow data. Two different Bayesian models are employed in order to account for the high level of uncertainty and error in hydrological model predictions [22, 23]. According to the first simple model, the hydrological parameters and the measurement uncertainty are unknown. The formulation is the nonlinear generalization of the inverse modeling with an unknown noise level that was discussed in Section 3.6.2. In addition to that, a second model is devised that considers also error correlation and model discrepancy as a function of time. The main behavior of the outflow is captured by the model, while a simple trend function accounts for the discrepancy. This more advanced representation of simulator discrepancy was already highlighted in Section 3.6.3.

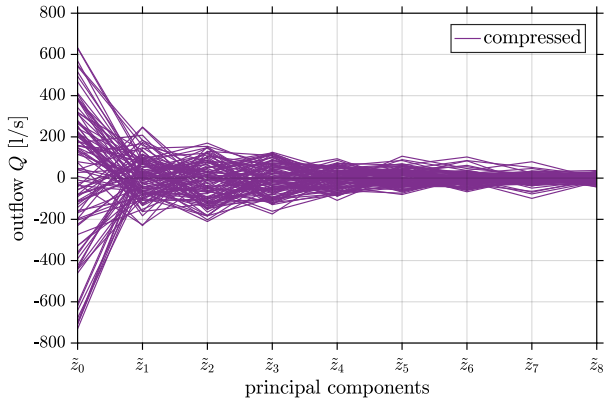


Figure 10.3: Principal component analysis.

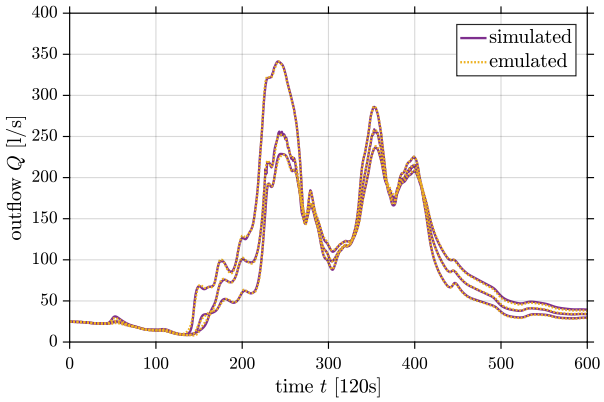


Figure 10.4: Final metamodel predictions.

The measurement data $\boldsymbol{y} = (y_0, \dots, y_{600})^\top$ comprise the observations of the outflow $y_i = Q(t_i)$ for $i = 0, \dots, 600$. They are represented as responses $\tilde{\boldsymbol{y}} = \mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x})$ of the forward model in Eq. (10.1) contaminated by random measurement noise and systematic model discrepancy. The unknown parameters of the forward and the error model can then be identified in a statistical data analysis.

10.3.1 Independent random errors

A first simple model takes only random measurement errors into account. They are assumed to act additively and independently on the forward model predictions. Thus the measured data are represented as

$$\boldsymbol{y} = \mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x}) + \boldsymbol{\varepsilon}. \tag{10.17}$$

Here, $\boldsymbol{\varepsilon}$ is a realization of a random vector with a Gaussian distribution $\pi(\boldsymbol{\varepsilon}|\sigma) = \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \sigma^2\boldsymbol{I})$, where the noise level $\sigma > 0$ is unknown. Consequently the following statistical model arises

$$\pi_1(\boldsymbol{y}|\boldsymbol{x}, \sigma) = \mathcal{N}(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x}), \sigma^2\boldsymbol{I}). \tag{10.18}$$

The likelihood function is simply $\mathcal{L}_1(\boldsymbol{x}, \sigma) = \mathcal{N}(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x}), \sigma^2\boldsymbol{I})$. Instead of merely maximizing the likelihood, a fully Bayesian approach is pursued. For any given prior distribution $\pi_1(\boldsymbol{x}, \sigma)$, the corresponding posterior is

$$\pi_1(\boldsymbol{x}, \sigma|\boldsymbol{y}) \propto \mathcal{L}_1(\boldsymbol{x}, \sigma)\pi_1(\boldsymbol{x}, \sigma). \tag{10.19}$$

In order to complete the setup, we specify a joint prior of the unknowns with the product structure $\pi_1(\boldsymbol{x}, \sigma) = \pi_1(\boldsymbol{x})\pi_1(\sigma)$ and $\pi_1(\boldsymbol{x}) = \pi_1(x_1) \dots \pi_1(x_8)$. The priors for the hydrological parameters $x_i \in [\underline{x}_i, \overline{x}_i]$ are normal distributions $\pi_1(x_i) = \mathcal{N}(x_i|\mu_{x_i}, \sigma^2_{x_i}, \underline{x}_i, \overline{x}_i)$ truncated at the respective parameter bounds $\underline{x}_i$ and $\overline{x}_i$. Before the truncation, the distributions are centered around the midpoint $\mu_{x_i} = (\underline{x}_i + \overline{x}_i)/2$ and their standard deviations $\sigma_i = (\overline{x}_i - \underline{x}_i)/6$ are set to the sixth part of the admissible range. Note that the prior for the hydrological parameters is different from the uniform distribution that the experimental design was sampled from. A uniform distribution $\pi_1(\sigma) = \mathcal{U}(\sigma|\underline{\sigma}, \overline{\sigma})$ with $\underline{\sigma} = 0 \times \mathrm{l/s}$ and $\overline{\sigma} = 100 \times \mathrm{l/s}$ is selected as the prior for the unknown noise level σ. The lower bound emerges naturally, whereas the upper bound is chosen so that it is highly probable that the true or best value is really contained in the supported interval.

10.3.2 Systematic model discrepancy

The second model is more sophisticated in that it also acknowledges other sources of uncertainty and error. In particular, model discrepancy and random error correlation are captured. We start the discussion by representing the measurement data as

$$\boldsymbol{y} = \mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x}) + \boldsymbol{\delta}(\boldsymbol{b}) + \boldsymbol{\varepsilon}. \tag{10.20}$$

This is the sum of the model response $\mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x})$ at the true $\boldsymbol{x}$ and two other terms that allow for a refined treatment of discrepancy and noise. The systematic modeling errors are absorbed into the term $\boldsymbol{\delta}(\boldsymbol{b})$, whereas $\boldsymbol{\varepsilon}$ captures the noise. We assume that the discrepancy is an unknown function of time that can be sufficiently well represented as

$$\delta(\boldsymbol{b}, t) = \sum_{\alpha=0}^{p} b_\alpha \Psi_\alpha(t). \tag{10.21}$$

Here, $\{\Psi_\alpha(t)\}_{\alpha=0}^p$ is a function basis with $P = p+1$ elements and $\boldsymbol{b} = (b_0, \dots, b_p)^\top$ denotes the unknown coefficients. The values $\delta_i(\boldsymbol{b}) = \delta(\boldsymbol{b}, t_i)$ of the discrepancy function at the measurement instances t_i for $i = 0, \dots, 600$ generate the discrepancy vector $\boldsymbol{\delta}(\boldsymbol{b}) = (\delta_0(\boldsymbol{b}), \dots, \delta_{600}(\boldsymbol{b}))^\top$.

The term $\boldsymbol{\varepsilon}$ is a realization of a random vector following a multivariate Gaussian distribution $\pi(\boldsymbol{\varepsilon}|\sigma,\tau) = \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \boldsymbol{\Sigma}(\sigma,\tau))$ with an unknown covariance matrix $\boldsymbol{\Sigma}(\sigma,\tau)$. For $i, j = 0, \dots, 600$ the entries of the covariance matrix are represented as

$$\Sigma_{i,j}(\sigma,\tau) = \sigma^2 \exp\left(-\frac{|t_i - t_j|}{\tau}\right). \tag{10.22}$$

As before, the standard deviation σ determines the noise level. The additionally introduced correlation length τ establishes a characteristic time scale of the error correlation. Both parameters σ and τ describing the covariance structure of the error process are unknown. In total, we have established the probabilistic data model

$$\pi_2(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau) = \mathcal{N}(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x}) + \boldsymbol{\delta}(\boldsymbol{b}), \boldsymbol{\Sigma}(\sigma,\tau)). \tag{10.23}$$

The likelihood function $\mathcal{L}_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau) = \mathcal{N}(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x}) + \boldsymbol{\delta}(\boldsymbol{b}), \boldsymbol{\Sigma}(\sigma,\tau))$ arises as a result. If one has a joint prior $\pi_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau)$, one obtains the posterior distribution by

$$\pi_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau|\boldsymbol{y}) \propto \mathcal{L}_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau)\pi_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau). \tag{10.24}$$

Some prior specifications are now overdue. We impose a joint prior distribution with the block-wise independence structure $\pi_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau) = \pi_2(\boldsymbol{x})\pi_2(\boldsymbol{b})\pi_2(\sigma)\pi_2(\tau)$. While the priors $\pi_2(\boldsymbol{x}) = \pi_1(\boldsymbol{x})$ and $\pi_2(\sigma) = \pi_1(\sigma)$ are not altered, we only have to set $\pi_2(\boldsymbol{b})$ and $\pi_2(\tau)$. The latter is chosen as $\pi_2(\tau) = \mathcal{U}(\tau|\underline{\tau}, \overline{\tau})$ with the lower bound $\underline{\tau} = 0 \times 120\,\mathrm{s}$ and a conservatively high upper bound $\overline{\tau} = 100 \times 120\,\mathrm{s}$.

We believe that the discrepancy $\delta(\boldsymbol{b}, t)$ is a rather smooth function of time. It is thus expanded in terms of the first normalized Legendre polynomials $\{\Psi_\alpha(t)\}_{\alpha=0}^p$ up to rather low degree $p = 5$. In fact there is no need to be picky while choosing the polynomial family here. Since the expansion coefficients $\boldsymbol{b}$ are mere tuning parameters which do not correspond to physically interpretable quantities, the specification of the prior $\pi_2(\boldsymbol{b}) = \pi_2(b_0) \dots \pi_2(b_5)$ is a bit delicate. We opt for Laplace distributions $\pi_2(b_i) = \mathrm{Laplace}(b_i|\mu_{x_i}, s_{x_i}) = (2s_{x_i})^{-1} \exp(-|b_i - \mu_{x_i}|/s_{x_i})$ for all $i = 0, \dots, 5$. They peak at the mean $\mu_{x_i} = 0$ and have the scale parameter $s_{x_i} = 10$ which leads to a standard deviation $\sigma_{x_i} = s_{x_i}\sqrt{2} \approx 15$.

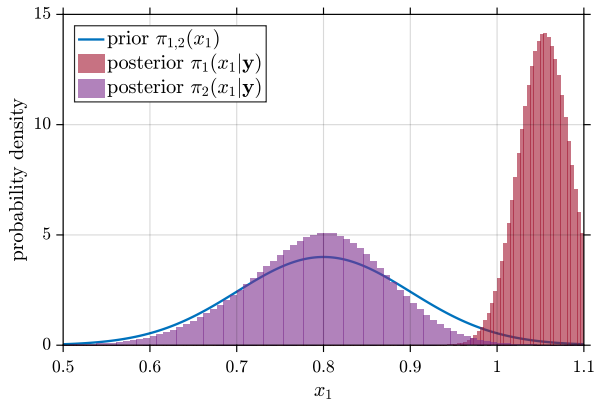
The double-exponential density $\mathrm{Laplace}(b_i|\mu_{x_i}, s_{x_i})$ decays exponentially with the absolute difference from the mean, whereas the bell-shaped density $\mathcal{N}(b_i|\mu_{x_i}, \sigma_{x_i}^2) = (2\pi\sigma_{x_i}^2)^{-1/2} \exp(-(b_i - \mu_{x_i})^2/(2\sigma_{x_i}^2))$ dies down with the squared difference. Accordingly, the Laplace distribution has a spikier peak and fatter tails than the Gaussian at the same time. Both sparsity of the coefficient vector and robustness with respect to the prior choice are promoted that way. While sparsity is not our main concern at this point, robustness can be indeed adduced as an argument for the Laplace prior. The specification of the scale parameter, however, remains more or less arbitrary after all.

10.3.3 Posterior distributions

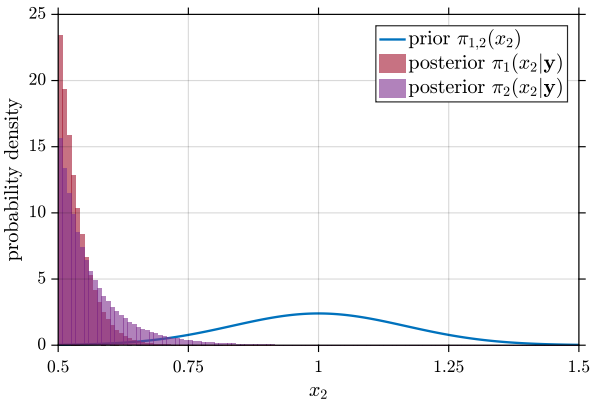
Now we perform fully Bayesian analyses by computing the two posterior distributions $\pi_1(\boldsymbol{x}, \sigma|\boldsymbol{y})$ and $\pi_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau|\boldsymbol{y})$ by means of MCMC sampling. The obtained surrogate model $\hat{\mathcal{M}}_p(\boldsymbol{x})$ is used in place of the original simulator $\mathcal{M}_{\boldsymbol{d}}(\boldsymbol{x})$ throughout the analyses. A random walk Metropolis algorithm with a Gaussian proposal distribution is deployed. Thirty parallel chains with 10^6 MCMC iterations are run for both Bayesian models. For the first model the parameters $(\boldsymbol{x}, \sigma)$ are updated altogether, while for the second model $\boldsymbol{x}$ and $(\sigma, \tau, \boldsymbol{b})$ are updated in two separate blocks. Roughly speaking, the posterior computations take half a day for the simple and about a week for the more complex model. The non-diagonal covariance matrix and the block-wise MCMC updates for the second model are responsible for the runtime difference.

First of all, we discuss the posterior marginals of the uncertain hydrological parameters $\boldsymbol{x}$. For $i = 1, \dots, 8$ the marginals $\pi_1(x_i|\boldsymbol{y})$ and $\pi_2(x_i|\boldsymbol{y})$ of both Bayesian models are shown in Fig. 10.5. Some marginals of the simple model feature posterior modes close to their bounds, i.e. see Figs. 10.5(a), 10.5(c) and 10.5(d) where the posteriors of x_1, x_3 and x_4 are depicted. Other marginals peak directly at the parameter bounds, i.e. the marginals of x_2, x_5 and x_6 in Figs. 10.5(b), 10.5(e) and 10.5(f), respectively. The posterior of x_7 in Fig. 10.5(g) is hardly different from the prior. A more complex structure is found in the marginal of x_8 that is shown in Fig. 10.5(h). It has two modes, one of which peaks at the lower parameter bound.

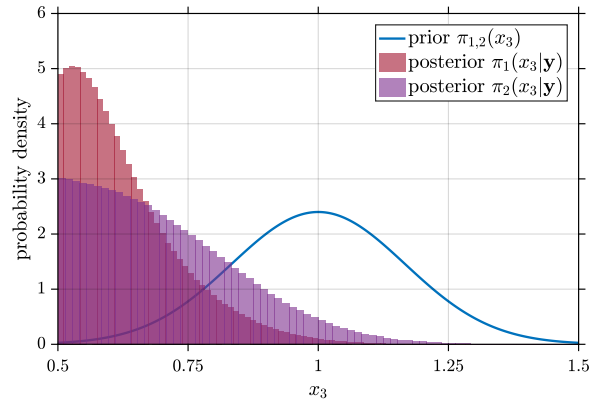
As compared to the simple model, the marginal posteriors of the second model with the discrepancy term are generally flattened out and shifted towards the prior means. Neither of the two models give clear evidence about the hydrological parameters. Relatively little information is gained through the Bayesian update. In a way, that the posteriors peak at the bounds even suggests that the problem is mis-specified.



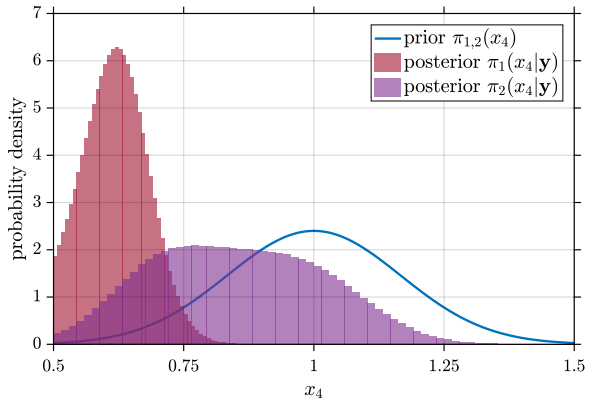
(a) Model parameter x_1.



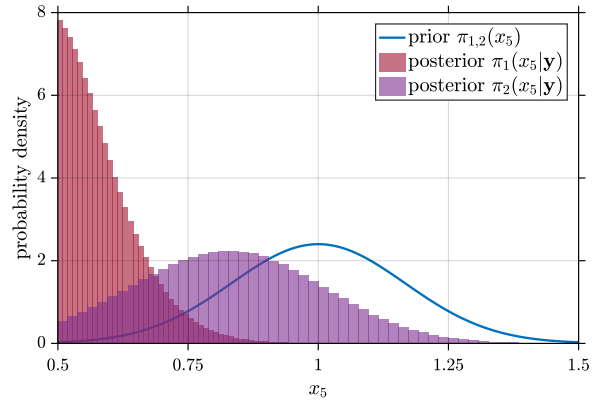
(b) Model parameter x_2.



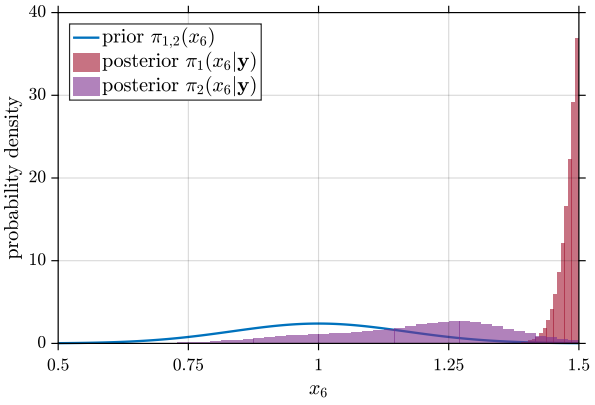
(c) Model parameter x_3.



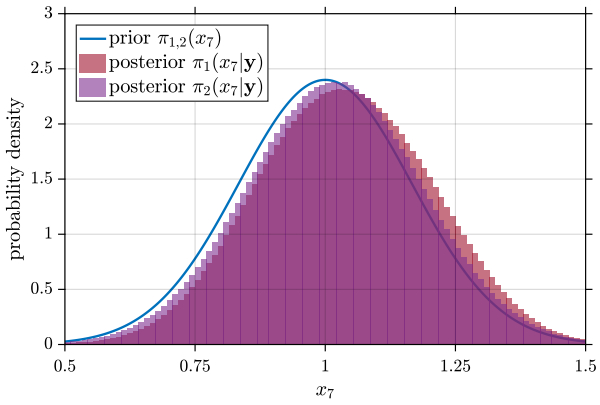
(d) Model parameter x_4.



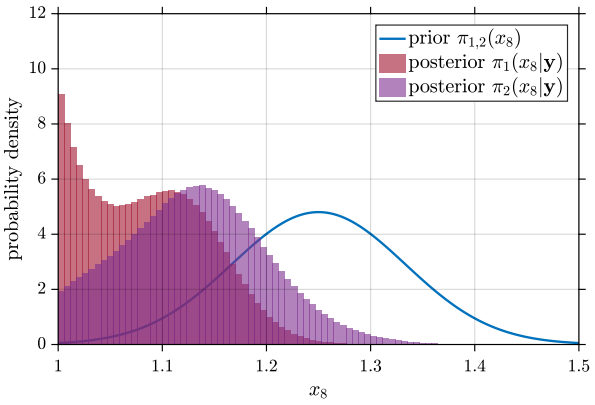
(e) Model parameter x_5.



(f) Model parameter x_6.



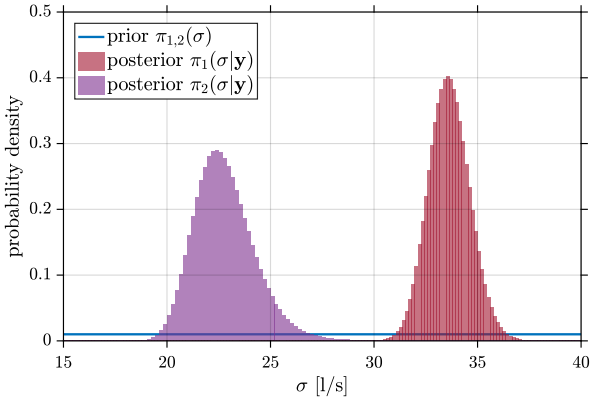
(g) Model parameter x_7.



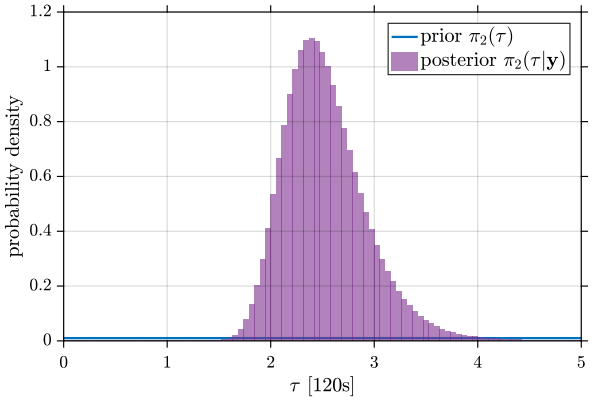
(h) Model parameter x_8.

Figure 10.5: Posterior marginals for the hydrological model.

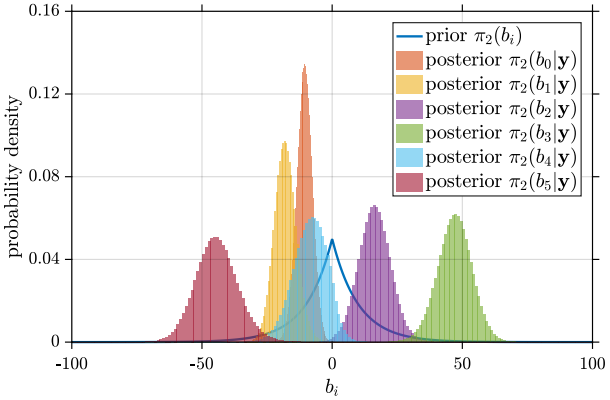
The posterior marginals of the parameters describing the random error model are shown in Fig. 10.6. As it can be seen from Fig. 10.6(a), the marginal $\pi_1(\sigma|\boldsymbol{y})$ suggests a higher value of the standard deviation σ than the marginal $\pi_2(\sigma|\boldsymbol{y})$. The reason is that according to the first model all errors are attributed to independent noise only. In the second model, those errors are also captured by the error correlation and model discrepancy. The marginal $\pi_2(\tau|\boldsymbol{y})$ of the correlation length τ is plotted in Fig. 10.6(b). It concentrates around a surprisingly low value. We speculate that the introduction and estimation of the discrepancy term effectively decorrelates the remaining sources of random error, which would explain this observation. In Fig. 10.6(c) all marginals $\pi_2(b_i|\boldsymbol{y})$ of the coefficients b_i with $i = 0, \ldots, 5$ are shown. Their actual units are discarded for the sake of simplicity. It is interesting to note that the parameters $\boldsymbol{b}$ of the discrepancy function are estimated quite clearly. Especially the constant and the linear term with their coefficients b_0 and b_1 have pronounced posterior shapes.



(a) Noise level σ.



(b) Correlation length τ.



(c) Discrepancy coefficients b_i.

Figure 10.6: Posterior marginals for the error model.

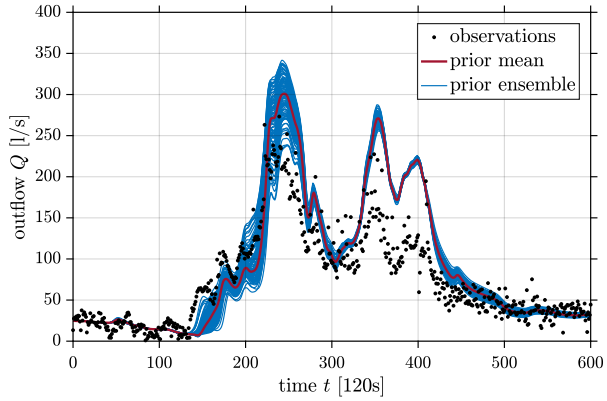
Some summaries of the posterior distributions $\pi_1(\boldsymbol{x}, \sigma|\boldsymbol{y})$ and $\pi_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau|\boldsymbol{y})$ are compiled in Table 10.3. These are point estimates of the unknown parameters, e.g. the posterior mean vectors $(\hat{\boldsymbol{x}}, \hat{\sigma}) = \mathbb{E}[\boldsymbol{x}, \sigma|\boldsymbol{y}]$ and $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}, \hat{\sigma}, \hat{\tau}) = \mathbb{E}[\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau|\boldsymbol{y}]$. Quantities whose dimension does not equal one are expressed in comparison to the units that were previously adopted. Posteriors that peak at the prior bounds are not summarized well by their mean values only. Therefore the modes $(\hat{\boldsymbol{x}}, \hat{\sigma})_{\text{MAP}} = \arg\max_{\boldsymbol{x},\sigma} \pi_1(\boldsymbol{x}, \sigma|\boldsymbol{y})$ and $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}, \hat{\sigma}, \hat{\tau})_{\text{MAP}} = \arg\max_{\boldsymbol{x},\boldsymbol{b},\sigma,\tau} \pi_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau|\boldsymbol{y})$ of the joint posterior densities are shown, too. They have been obtained through maximizing the logarithms of the unnormalized posterior densities, i.e. the log–likelihood function plus the log–prior density. Note that the individual components of the joint posterior density mode do not have to coincide with the maxima of the marginal densities.

Table 10.3: Posterior summaries.

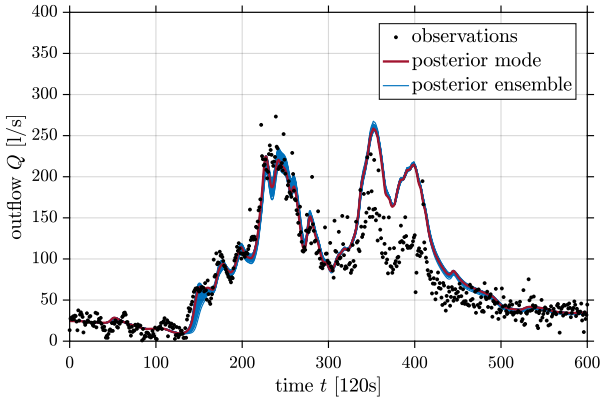
		$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	$\hat{x}_4$	$\hat{x}_5$	$\hat{x}_6$	$\hat{x}_7$	$\hat{x}_8$	$\hat{\sigma}$	$\hat{\tau}$	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$
$\pi_1(\cdot\|\boldsymbol{y})$	Mean	1.05	0.54	0.63	0.62	0.59	1.48	1.04	1.09	33.63	-	-	-	-	-	-	-
	Mode	1.06	0.50	0.55	0.62	0.50	1.49	0.99	1.00	33.04	-	-	-	-	-	-	-
$\pi_2(\cdot\|\boldsymbol{y})$	Mean	0.79	0.56	0.70	0.85	0.84	1.18	1.02	1.13	22.78	2.53	−10.63	−17.97	16.33	47.05	−8.38	−44.31
	Mode	0.71	0.50	0.50	0.71	0.59	0.91	1.01	1.03	19.95	1.80	−13.33	−20.25	19.39	46.70	−10.25	−42.72

After having explored the posterior distribution, one can check the obtained results for consistency by comparing an ensemble of prior and posterior predictions with the data. We start by comparing the posterior $\pi_1(\boldsymbol{x}, \sigma|\boldsymbol{y})$ of the first model with the correspondent prior $\pi_1(\boldsymbol{x}, \sigma)$ in this regard. See Fig. 10.7 for that purpose. In Fig. 10.7(a) the forecasts of the outflow are shown for one hundred input values that were randomly sampled from the prior. Likewise Fig. 10.7(b) shows the predictions for the same number of posterior samples that were obtained from the MCMC chains by an appropriate thinning. Moreover, the time trajectory for the posterior mode is highlighted. The measurement uncertainty is not accounted for in those figures. As it can be seen, the prediction ensemble for the prior contains more uncertainty than for the posterior.

The adjustment of the model parameters associated with the Bayesian update does not significantly reduce the systematic discrepancy between the simulated and the measured outflows from the drainage basin. The underlying reason is that varying the input parameters of the hydrological simulator and the level of independent noise does not allow for establishing full consistency between the simulations and the observations, especially in the second half of the covered time interval. This was already clear after the discussion of Fig. 10.2(b) and actually led to the inclusion of a correlation and discrepancy term in the second model.



(a) Prior predictions.



(b) Posterior predictions.

Figure 10.7: Stochastic model predictions.

We now investigate how well the posterior mode of $\pi_2(\boldsymbol{x}, \boldsymbol{b}, \sigma, \tau|\boldsymbol{y})$ aligns with the data. The mode estimate of the discrepancy function $\hat{\delta}(t) = \delta(\hat{\boldsymbol{b}}, t)$ is plotted in Fig. 10.8. It indicates a trend that the model underpredicts the actual rainfall in roughly the interval $t/120\,\mathrm{s} \in [100, 250]$ and overpredicts in $t/120\,\mathrm{s} \in [250, 500]$. These mis-predictions occur more or less for the period $t/120\,\mathrm{s} \in [100, 450]$ of the precipitation event that was shown in Fig. 10.2(a). At the boundaries, say for $t/120\,\mathrm{s} \in [0, 100]$ and $t/120\,\mathrm{s} \in [500, 600]$, the discrepancy vanishes as far as the low-degree polynomial representation admits. The accordingly corrected predictions $\hat{\mathcal{M}}_p(\hat{\boldsymbol{x}}) + \hat{\boldsymbol{\delta}}$ are depicted in Fig. 10.9. They align with the data reasonably well. One, two and three $\hat{\sigma}$ prediction intervals are added so as to visualize the posterior mode prediction uncertainty. Due to the additive and symmetric error model, the intervals extend to negative outflow values. Since these values are physically nonsensical, they shall be ignored.

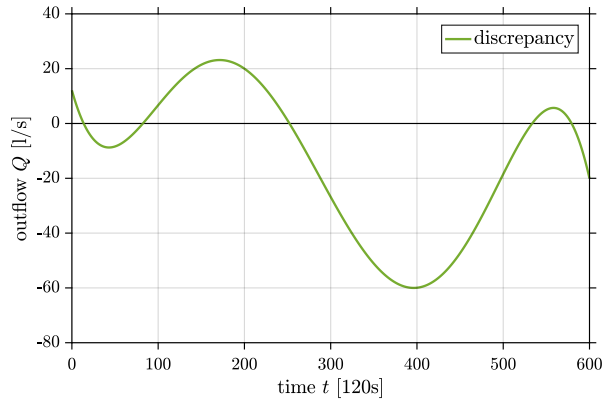


Figure 10.8: Model discrepancy.

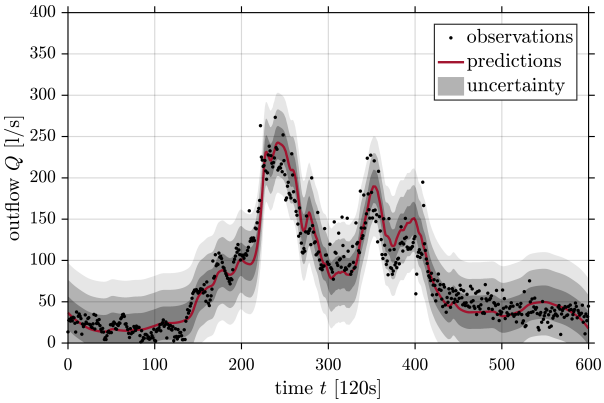


Figure 10.9: Corrected predictions.

10.4 Discussion and conclusion

Probabilistic calibration of a hydrological urban drainage simulator was accomplished in this chapter. A combination of techniques for Bayesian and surrogate modeling was deployed for that purpose. The original training runs of the simulator were statistically compressed and subsequently translated into a functional emulator. Inference was then based on the exploration of the posterior distributions related to two different Bayesian models through Markov chain Monte Carlo.

The two models differ in their ability to capture measurement and modeling uncertainties. With the first model, that only acknowledges independently varying errors, the hydrological parameters were calibrated. With the second Bayesian model, that also includes random error correlation and systematic model discrepancy, we could additionally quantify the mismatch between the model predictions and the data throughout the rainfall event. This was, however, accompanied by a lack of interpretability with regard to the corresponding estimates of the hydrological parameters.

This last observation was already foreshadowed in the very setup of the problem. While the catchment area has multiple outlets, only a single one at the wastewater treatment plant was considered for parameter calibration purposes. A coarse-grained parametrization of the unknowns based on crude averages over hundreds of sub-catchments and channels was used. Moreover, the whole procedure was dependent on a single precipitation event for which the rainfall record was taken as if it were measured without error. In this context one also has to mention that the predictions of the forward model were highly uncertain and only weakly sensitive to the calibration parameters.

The preceding discussion suggests a number of possible extensions and future improvements that would allow for more complex and realistic modeling. A more refined representation of the uncertain hydrological parameters could be based on a finer graining of the spatial resolution. Sparse polynomial chaos expansions and advanced stochastic sampling schemes would allow one to cope with the associated increase in dimensionality. Errors and uncertainties in the rainfall input data could be considered by treating and inferring the rainfall as additional unknowns. If the data for various different precipitation events were available, a more thorough representation and management of the encountered uncertainties could be based on hierarchical Bayesian modeling as developed in Chapters 4 and 5.

References

[1] K. Beven. *Rainfall-Runoff Modelling: The Primer*. 2nd ed. Chichester, West Sussex, UK: Wiley-Blackwell, 2012. DOI: 10.1002/9781119951001.

[2] D. Machac. "Mechanistic Emulators as Surrogates to Slow Hydrological Models". PhD thesis. Zürich, Switzerland: Swiss Federal Institute of Technology (ETH Zürich), 2015.

[3] V. P. Singh and D. A. Woolhiser. "Mathematical Modeling of Watershed Hydrology". In: *Journal of Hydrologic Engineering* 7.4 (2002), pp. 270–292. DOI: 10.1061/(ASCE)1084-0699(2002)7:4(270).

[4] E. Todini. "Hydrological catchment modelling: past, present and future". In: *Hydrology and Earth System Sciences* 11.1 (2007), pp. 468–482. DOI: 10.5194/hess-11-468-2007.

[5] E. Todini. "History and perspectives of hydrological catchment modelling". In: *Hydrology Research* 42.2–3 (2011), pp. 73–85. DOI: 10.2166/nh.2011.096.

[6] L. A. Rossman. *Storm Water Management Model: User's Manual*. Version 5.1. US EPA Office of Research and Development. Washington, D.C., USA, 2015.

[7] P. Reichert, G. White, M. J. Bayarri, and E. B. Pitman. "Mechanism-based emulation of dynamic simulation models: Concept and application in hydrology". In: *Computational Statistics & Data Analysis* 55.4 (2011), pp. 1638–1655. DOI: 10.1016/j.csda.2010.10.011.

[8] C. Albert. "A mechanistic dynamic emulator". In: *Nonlinear Analysis: Real World Applications* 13.6 (2012), pp. 2747–2754. DOI: 10.1016/j.nonrwa.2012.04.003.

[9] D. Machac, P. Reichert, and C. Albert. "Emulation of dynamic simulators with application to hydrology". In: *Journal of Computational Physics* 313 (2016), pp. 352–366. DOI: 10.1016/j.jcp.2016.02.046.

[10] D. Machac, P. Reichert, J. Rieckermann, and C. Albert. "Fast mechanism-based emulator of a slow urban hydrodynamic drainage simulator". In: *Environmental Modelling & Software* 78 (2016), pp. 54–67. DOI: 10.1016/j.envsoft.2015.12.007.

[11] M. D. McKay, R. J. Beckman, and W. J. Conover. “A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code”. In: *Technometrics* 21.2 (1979), pp. 239–245. DOI: 10.2307/1268522.

[12] G. H. Dunteman. *Principal Components Analysis*. Quantitative Applications in the Social Sciences 69. Newbury Park, California, USA: Sage Publications, Inc., 1989. DOI: 10.4135/9781412985475.

[13] J. E. Jackson. *A User’s Guide to Principal Components*. Wiley Series in Probability and Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 1991. DOI: 10.1002/0471725331.

[14] I. T. Jolliffe. *Principal Component Analysis*. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag, 2002. DOI: 10.1007/b98835.

[15] G. Blatman and B. Sudret. “Principal component analysis and Least Angle Regression in spectral stochastic finite element analysis”. In: *11th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP11)*. Ed. by M. H. Faber, J. Köhler, and K. Nishijima. Leiden, Netherlands: CRC Press/Balkema, 2011. Chap. 82, pp. 669–676. DOI: 10.1201/b11332-101.

[16] G. Blatman and B. Sudret. “Sparse polynomial chaos expansions of vector-valued response quantities”. In: *11th International Conference on Structural Safety and Reliability (ICOSSAR 2013)*. Ed. by G. Deodatis, B. R. Ellingwood, and D. M. Frangopol. Leiden, Netherlands: CRC Press/Balkema, 2013. Chap. 434, pp. 3245–3252. DOI: 10.1201/b16387-469.

[17] M. Loève. *Probability Theory*. 4th ed. 2 vols. Graduate Texts in Mathematics 45–46. New York: Springer-Verlag, 1977–1978.

[18] D. Vidaurre, C. Bielza, and P. Larrañaga. “A Survey of L_1 Regression”. In: *International Statistical Review* 81.3 (2013), pp. 361–387. DOI: 10.1111/insr.12023.

[19] H. Zhang and R. H. Zamar. “Least angle regression for model selection”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 6.2 (2014), pp. 116–123. DOI: 10.1002/wics.1288.

[20] G. Blatman and B. Sudret. “Adaptive sparse polynomial chaos expansion based on least angle regression”. In: *Journal of Computational Physics* 230.6 (2011), pp. 2345–2367. DOI: 10.1016/j.jcp.2010.12.021.

[21] S. Marelli and B. Sudret. *UQLab User Manual: Polynomial Chaos Expansions*. Version 0.9-104. Chair of Risk, Safety & Uncertainty Quantification, ETH Zürich. 2015.

[22] D. Del Giudice, M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann. “Improving uncertainty estimation in urban hydrological modeling by statistically describing bias”. In: *Hydrology and Earth System Sciences* 17.10 (2013), pp. 4209–4225. DOI: 10.5194/hess-17-4209-2013.

[23] D. Del Giudice, P. Reichert, V. Bareš, C. Albert, and J. Rieckermann. “Model bias and complexity – Understanding the effects of structural deficits and input errors on runoff predictions”. In: *Environmental Modelling & Software* 64 (2015), pp. 205–214. DOI: 10.1016/j.envsoft.2014.11.006.

Chapter 11

Conclusion

After the reading it is now time to retrospect and recapitulate. The starting points of this dissertation were two quite broad research questions. First of all, how can we cope with both epistemic uncertainty and aleatory variability in Bayesian inverse problems? Second of all, how can we overcome the limitations of sampling-based methods for computing the posterior probability distribution? The provided answers and some of the ensuing issues are summarized and discussed in these concluding remarks.

11.1 Hierarchical modeling

In reply to the first guiding research question, a hierarchical formulation and integrative solution of Bayesian inverse problems under uncertainty and variability was presented in Chapter 4. It allows one to perform uncertainty quantification and data analysis in complex experimental situations, where a forward model predicts the observable quantities, but the inputs are uncertain or variable. Different types of forward model inputs were distinguished in this respect. There are global model parameters that are constant throughout and unknown. Further unknowns are variable quantities that take on different values during the experimentation. Their distributions are determined by hyperparameters that are often unknown, or else, they are already well-known or even controllable. The epistemic uncertainty of the global parameters can be reduced within the developed multilevel framework and one can infer the population distributions of the variable quantities. An optimal combination of the information available from models, experiments and experts is achieved that way. On this basis, various aspects of ensemble heterogeneity and temporal or spatial variation can be studied.

Since Bayesian inverse problems in the presence of uncertainties are often strongly simplified and their solutions are occasionally misunderstood, it is believed that this work is of high relevance and value. Variable quantities are indeed often mistreated as constants in current practice and a popular fallacy is to misinterpret subjective Bayesian measures of uncertainty, especially posterior probabilities, as objective frequencies. The hierarchical framework establishes a solid foundation for inverse problems under uncertainty and thereby clarifies these matters. Its potency and flexibility were demonstrated in a number of realistic engineering applications. It will continue to be used in the future.

11.2 Hamiltonian Monte Carlo

As it was discussed, Bayesian inversion in the presence of multiple sources and types of uncertainty poses considerable computational challenges. The high-dimensionality of the parameter space causes difficulties for traditional Markov chain Monte Carlo sampling techniques, while a lower-dimensional but yet mostly equivalent reformulation calls for costly evaluations of an integrated likelihood function instead. Hamiltonian Monte Carlo was proposed as an efficient sampling algorithm in Chapter 5. This is a gradient-driven sampler with ancillary parameters which is inspired by systems from classical physics. Here the posterior is explored through a point mass moving in a potential well that is proportional to minus its log-density. It was shown that this updating scheme is ideally suited for the high-dimensional spaces arising in hierarchical inverse problems. The posterior was sampled almost independently and a simple Metropolis–Hastings algorithm was easily outperformed.

Future research efforts will involve the employment of polynomial chaos expansions in hierarchical models. After the specification of an appropriate weight function for the model input parameters and the assignment of the associated orthogonal polynomials as basis functions, accordingly constructed metamodels will accelerate the computations. In conjunction with Hamiltonian Monte Carlo sampling, they could also assist in finding the

necessary derivatives of the forward model and the posterior log-density, which would offer an alternative to adjoint modeling, automatic differentiation and finite differencing. This idea is actually not limited to hierarchical models only, it can be used in non-hierarchical problems just as well. In turn, this raises the question of how accurate the derivatives of a polynomial approximation of the forward model are.

11.3 Realistic applications

The developed multilevel framework was used for solving the identification problem of the NASA Langley challenge in Chapter 6. Here the goal was to calibrate an unmanned aircraft model subjected to adverse conditions such as structural damage or component failure. A physical model, experimental data and statistical information regarding the uncertainty and variability of the relevant quantities were provided by the challenge organizers. This was translated into a Bayesian hierarchical model whose parameters are related to aerodynamic conditions and the loss of control effectiveness. An oddity of the challenge consisted in the fact that the forward model is perfect in the sense that the data are noise-free. As a consequence, the likelihood function had to be constructed as the solution to a subsidiary uncertainty propagation problem. The latter could be addressed by Monte Carlo simulation and kernel density estimation, which was accompanied by a deformation of the corresponding posterior distribution. Even though it was tried to investigate and moderate this undesirable side effect based on heuristic checks and partial data augmentation, the rigorous analysis of the induced posterior approximations remains an important issue for the future.

Hierarchical Bayesian modeling was also employed for assessing masonry wall compressive strengths in Chapter 7. The most important property of structural masonry is the compressive strength perpendicular to the bed joints. Statistically predicting this key characteristic of masonry walls based on tests of brick units and mortar samples, that belong to the same population used in the construction of the wall, was the objective of this study. Previous efforts in that direction fail in providing satisfactory predictions and quantifying the inevitable uncertainties. A probabilistic model based on lognormal distributions with unknown hyperparameters was constructed. It was trained with full-scale tests of masonry walls and tests of the corresponding brick and mortar ensembles, that were executed by Dr. Nebojsa Mojsilovic at the Institute of Structural Engineering of ETH Zürich. The statistically predictive relationship that was obtained hereby could be validated by applying it to an independent test set. Its performance proved to be superior to previous attempts, which in the future, when more data will become available and more complex models can be created and calibrated, is even expected to improve. Adaptations of the approach taken will be useful for the investigation of many kinds of composite systems that are constructed of similar elements from certain populations.

Another application of Bayesian inference to a real-world engineering problem was considered in Chapter 10. The ambition was to calibrate a dynamic urban drainage simulator, not under parametric variability, but in the presence of various forms of modeling errors. Experimental data, runs of the hydrological simulator and a prior distribution were made available to that end by the Swiss Federal Institute of Aquatic Science and Technology. A difficulty was that the forward model was not provided in an executable form, such that it could not be run for arbitrary input values. Therefore, a response surface was fitted to the training runs at hand. This was done by reducing the large number of time-variant response variables by means of principal component analysis and metamodeling the lower number of principal components with sparse polynomial chaos expansions. In order to account for correlated random errors and systematic model discrepancy, sophisticated statistical models had to be deployed. The whole chain of analyses made model corrections possible and resulted in well-calibrated predictions. Beyond the hydrological case study performed, the proposed combination of methods will generally facilitate to deal with legacy code and data.

11.4 Novel methods

A novel method for computing the posterior distribution by means of spectral likelihood expansions was developed in Chapter 8. Spectral Bayesian inference tries to beat the convergence rate of Monte Carlo approaches, where samples are treated and processed locally, by exploiting global structures of the problem and regularity properties of the likelihood function, in particular its smoothness as measured by its differentiability. Based on an expansion of the likelihood in terms of polynomials that are orthogonal with respect to the prior weight, an orthogonal series representation of the joint posterior density was derived. The nonparametric expression was interpreted as a perturbation series around the prior, which eventually suggested an adaptive procedure based on a recurrent baseline density change. While the posterior marginals emerge as sub-expansions, the model evidence and the posterior moments are related to the expansion coefficients. Furthermore, posterior uncertainty propagation can be accomplished by prior polynomial chaos expansions. Classical distribution fitting

and an inverse heat conduction problem served as low-dimensional application examples for demonstrating and benchmarking this rather unconventional technique. One of the next steps will be to combine spectral Bayesian inference with variational strategies and Laplace approximations. A rough approximation of the posterior can be found first with one of those well-established techniques. The obtained approximation could then be used as an auxiliary expansion baseline and it would be appropriately corrected so as to approach the true posterior. In turn, this will require to construct polynomial chaos expansions with arbitrary input measures, which will then help in tackling higher-dimensional problems. Whether there are bases that are more beneficial to likelihood expansions than multivariate polynomials is an interesting open research question. The most important question might be how to assess the errors in the computed expansions coefficients and their impact on the actually relevant posterior moments and expectation values.

Another recently devised method for computational Bayesian inference was investigated in Chapter 9. It is based on the diligent construction of a deterministic coupling of distributions or a transformation of random variables. The basic idea is to find a transport map that morphs a prior-distributed random vector into a posterior-distributed one. Motivated by optimal transportation theory, an appropriate map can be found by solving an optimization problem featuring an information-theoretic optimality criterion. More specifically, the Kullback–Leibler divergence from the back-transformed posterior to the prior is minimized. This approach was implemented and discussed in the context of variational inference. The bottom line is that transformations of the prior establish a remarkably flexible class of candidate posteriors, but the actual computation of a transport map is expensive. It is envisaged to combine inferential mapping with Monte Carlo sampling. Indeed, one could transform some standard distribution, which is easy to sample from, into a distribution that mimics the posterior, which in turn would aid in assessing the exact posterior. By the same token, one could also transform the posterior in such a way that it resembles some familiar probability distribution.

Spectral likelihood expansions and optimal inferential maps strike radically different paths of computing the posterior distribution. Notwithstanding that they may not yet be as full-fledged as Markov chain theory, they have the potential to attain maturity in the future and to lay a new foundation for computational Bayesian inference. This would answer the second research question of how to remedy the major shortcomings and numerous inconveniences of Markov chain Monte Carlo sampling.

All in all, the made developments allow for a thorough and efficient data and uncertainty analysis. The framework for Bayesian inversion under multiple types of uncertainty enables the principled study of many complex systems. Moreover, the presented numerical approaches to Bayesian inference offer completely new possibilities of representing and characterizing the posterior. These approaches establish promising alternatives to conventional methods and they will hopefully stimulate many future developments.