Diss. ETH No. 17822

Infinite-Dimensional Exponential Families in Cluster Analysis of Structured Data

A dissertation submitted to
ETH Zurich

for the degree of
Doctor of Sciences

presented by
Peter Orbanz
Dipl.-Inf. (University of Bonn)
born 4. September 1975
citizen of Germany

accepted on the recommendation of
Joachim M. Buhmann, examiner
Zoubin Ghahramani, co-examiner
Sara van de Geer, co-examiner

2008

Abstract

The principal focus of the present thesis is the cluster analysis of structured data, in particular with spatial and temporal coupling structure, and with ordinal structure of individual observations. The thesis studies grouping models and algorithms which (i) take into account structure such as spatial and temporal smoothness and (ii) estimate or select the model order, or number of clusters, from the data. Applications include segmentation of image and remote sensing data, video sequences, and cluster analysis of rank data.

For the cluster analysis of ordinal data, we introduce a mixture model suitable for the simultaneous representation of partial rankings of different lengths. Unlike grouping models for ordinal data previously available in the literature, the model permits the analysis of heterogeneous data sets. For segment analysis of noisy image data, we show that nonparametric Bayesian mixture models can be combined with Markov random fields. The resulting class of models simultaneously estimates the model order and a segmentation of the image, under a smoothness constraint on the segment solution. Video sequences exhibit a temporal dependence structure, similar to spatial coupling in still images. A sequence model of conditional Dirichlet processes is proposed, based on the conjugate nature of the Dirichlet process, which estimates cluster structure evolving over time. By conditioning on a previous estimate, such methods perform *model order adaptation* rather than model order selection. Based on the observation that the Dirichlet process inherits a number of key properties from its finite-dimensional marginals, we study the general construction of infinite-dimensional ("nonparametric") Bayesian models. It is shown that infinite-dimensional conjugate models are generated as projective limits of finite-dimensional conjugate models, in particular of those with exponential family components. An extension theorem for conditional measures and general construction criteria are given, and sufficiency and conjugacy properties of finite-dimensional Bayesian models are shown to be preserved under extension to the infinite-dimensional case.

Zusammenfassung

Schwerpunkt der vorliegenden Arbeit ist die Gruppierungsanalyse (Cluster-Analyse) strukturierter Daten, insbesondere von Daten mit räumlicher und zeitlicher Kopplungsstruktur, sowie mit ordinaler Struktur einzelner Messungen. Die Dissertation studiert Gruppierungsmethoden welche (i) Strukturen wie räumliche und zeitliche Glattheit in die Analyse miteinbeziehen, und (ii) die *Modellordnung*, d.h. die Anzahl der Gruppen, aus den vorliegenden Daten schätzen. Anwendungen umfassen die Segmentierung von Bildern und Fernerkundungsdaten, von Videosequenzen, und die Gruppierungsanalyse ordinaler Präferenzdaten.

Zur Gruppierungsanalyse von Ordinaldaten wird ein Mixturmodell vorgestellt, welches die simultane Repräsentation partieller ordinaler Messungen verschiedener Länge ermöglicht, und damit die ganzheitliche Analyse heterogener Datensätze, welche mit zuvor in der Literatur verfügbaren Ansätzen nicht möglich ist. Zur Segmentierung verrauschter Bilder wird gezeigt, dass nichtparametrische Bayessche Mixturmodelle mit Markovschen Zufallsfeldern kombinierbar sind. Die resultierende Modellklasse erlaubt die simultane Schätzung der Modellordnung und der Segmente unter einer Glattheitsbedingung an die Segmentierungslösung. In Videosequenzen existiert, ähnlich der räumlichen Kopplungsstruktur in Einzelbildern, eine zeitliche Abhängigkeit. Es wird gezeigt, wie Dirichlet-Prozesse unter Ausnutzung ihrer konjugierten Eigenschaften zur Schätzung von über die Zeit evolvierender Gruppenstruktur eingesetzt werden können. Durch Konditionierung auf die jeweils vorangehende Lösung führen solche Verfahren eine *Adaption* (anstatt einer Selektion) der Modellordnung durch. Wir untersuchen die allgemeine Konstruktion unendlich-dimensionaler ("nichtparametrischer") Bayesscher Modelle, ausgehend von der Beobachtung, dass die Marginalverteilungen des Dirichlet-Prozesses Modelle in der Exponentialfamilie darstellen, und eine Reihe wichtiger Eigenschaften an den Prozess vererben. Es wird gezeigt, dass nichparametrische konjugierte Modelle als projektive Limiten solcher endlich-dimensionaler Modelle entstehen, welche ebenfalls konjugiert sind, und damit insbesondere von Modellen in der Exponentialfamilie. Wir geben einen Fortsetzungssatz für bedingte Verteilungen und allgemeine Konstruktionsbedingungen an, und zeigen, inwiefern Suffizienz- und Konjugiertheits-Eigenschaften endlich-dimensionaler Bayesscher Modelle auf den unendlich-dimensionalen Fall übertragbar sind.

Contents

Chapter 1

Introduction

Data clustering addresses the problem of partitioning a heterogeneous set of data into homogeneous groups, i. e. groups consisting of data values mutually similar according to a given mathematical measure of similarity (Jain *et al.*, 1999; Duda and Hart, 1973). If the number of groups or categories is not known beforehand, the problem of estimating it from data is referred to as *model order selection* (Stoica and Selen, 2004). The principal focus of the present thesis is to contribute to the development of a new generation of clustering methods, which (i) address the entire clustering problem, including model order selection, by a model-based approach, allow (ii) out-of-sample prediction, (iii) seamless integration with different types of constraints, and (iv) generic application to a wide variety of data.

Unsupervised learning methods reduce the complexity of data, either as a preprocessing step for further automatic methods, or to render large sets of complex data accessible to human analysis (see e. g. Bishop, 2006). Automatic methods can benefit from such reduction techniques in a number of ways. One example is the estimation of category-specific models, such as topic-specific specialization of language models, after the data has been partitioned into categories by a clustering algorithm. Others include dimension reduction and feature selection techniques which reduce the effective dimensionality of the problem. Data analysis by humans is first and foremost a visual endeavor – a one- or two-dimensional data set is typically analyzed by "taking a look", that is, by preparing a plot. If the data is of higher dimension, unsupervised learning provides a collection of dimension reduction techniques which attempt to reduce the data to a lower-dimensional surrogate data set that still contains the relevant structures or patterns of the input data. If the number of observations is large, clustering methods can provide both a subdivision of the data into meaningful categories, and a summary of each category in form of a statistical model of the cluster. In this sense, unsupervised learning techniques serve to map raw data of a format unaccessible to humans into the scope of human cognition.

Applications of clustering range from image segmentation in computer vision (Forsyth and Ponce, 2003), through topic modeling in natural language processing (Hofmann, 1999) and vector quantization in signal processing (Gersho and Gray, 1992), to a variety of problems in all areas of science and technology that produce massive amounts of measurement data, such as biotechnology or sensor data processing. For example, most natural language processing techniques are based on the estimation of a language model, i. e. a mathematical model for the probability of a word to occur in a given context. Significant gains in accuracy for such models have been observed when the model (or the applied smoothing techniques) are estimated specific to topics, i. e. to the use of language in the context of a given subject. Automatic identification of topics, as groups of keywords, in a collection of texts is a classic example of a clustering problem (Hofmann, 1999; Blei *et al.*, 2003).

Most applications of clustering in this thesis are motivated by mid-level computer vision – computer vision methods that combine the strictly local outputs of low-level vision into a global model of the image, but do not attempt to infer image semantics (Forsyth and Ponce, 2003). Clustering is used for *image segmentation*, the subdivision of an image into coherent regions. Given an input image, a segmentation algorithm first computes low-level descriptors. These are, in terms of machine learning, statistics or features computed locally on small image patches. Descriptors are chosen to measure similarity, i. e. to take similar values on similar patches. A segmentation can be obtained by clustering the feature values and interpreting each cluster as an image segment. This approach is widely applicable due to its modular structure, and to some degree separates the vision-specific problem (the feature extraction) from the machine learning problem (clustering).

1.1 Thesis Overview

The central subject of the present thesis are methods that integrate the model order selection problem into the generative model. Model order selection is typically performed by model switching rules or heuristics: Solutions with different number of clusters constitute different models. Each possible model order (number of clusters) defines one model class, and the number of clusters is chosen on a given data set by means a criterion for selecting one such class. Popular examples include complexity penalties, such as the Bayesian information criterion (see Stoica and Selen, 2004, for an overview), the stability method (Dudoit and Fridlyand, 2002; Lange *et al.*, 2004), or reversible jump sampling (Green, 1995). The resulting solutions are not comparable over different model classes, and not generative w. r. t. the number of clusters, that is, a probability can be assigned to a given data set only conditional on the number of categories. We build on existing work on clustering with Dirichlet process mixture models, which provide a generative framework for the model order selection problem (Ferguson, 1973; Blei, 2004).

Algorithmic contributions are motivated by computer vision, where the performance of clustering algorithms can be improved by additional constraints adapted to the special structure of visual data. We show how both types of constraints common in mid-level vision, spatial and temporal smoothness, seamlessly integrate with nonparametric Bayesian clustering in a principled manner. Theoretical contributions address the representation problem for Bayesian nonparametric methods: Unlike the parametric models commonly used in machine learning, Dirichlet process models are not representable in closed form as probability densities. This considerably complicates their generalization to larger classes of models, and obstructs the derivation of an explicit, cost-function based formulation of unsupervised learning problems. We derive a representation of nonparametric Bayesian models that can be regarded as a weak analogue of the density representation of parametric models, show under which conditions the posteriors of such nonparametric Bayesian models are analytically tractable, and for such models identify their sufficient statistics in an effort to provide a nonparametric Bayesian analogue of a closed-form treatment.

1.1.1 Methods

The models and algorithms presented and discussed here draw on a number of techniques from mathematical statistics and algorithmic data analysis, the most important of which are briefly outlined below.

Nonparametric statistics. In the terminology of statistics, a probability model is a class of probability distributions, and the model is fitted to data by selecting the element of the class which best accounts for the observations. Such a model is called *parametric* if its elements are distinguished from one another by the values of a set of parameters, and if the number of model parameters does not depend on the number of observations. The fineprint about a fixed number of parameters prevents the model from growing more complex as the number of observations increases. A *nonparametric* model is also a parameterized probability model, but one which does not match the fineprint, i. e. the number of model parameters may grow with respect to sample size (Wasserman, 2006). A density estimation model that fits a single Gaussian bell-curve to a given data set, and adjusts the parameter estimate as more data comes in, is a parametric model. A Parzen density estimator, which smoothes the data by centering one Gaussian at each data point, is nonparametric. The fundamental trade-off between parametric and nonparametric models is that parametric models tend to come with more theoretical guarantees and faster convergence rates (how much data is needed to choose the optimal element of the model class), whereas nonparametric models are better suited for problems requiring adaptation, and often work impressively well in practice.

In clustering, representing each group by a parametric model is usually a reasonable assumption, reflecting the intuition that individual clusters should be of sufficiently simple structure. If the number of clusters is fixed, such parametric components can be combined to form an overall parametric model of the data (McLachlan and Peel, 2000). But if the number of components is to be estimated from the data, and can change from one data set to another, then the overall model must be nonparametric. The work presented in this thesis is primarily concerned with nonparametric models. A further fundamental choice in the design of probabilistic models is whether or not the model parameters should be treated as random quantities. Regarding parameters as random variables results in a *Bayesian* model, in which the parameters have a probability distribution, and statistical estimation attempts to determine the parameter's distribution conditional on a data set. Combination of the Bayesian and nonparametric approach has long been fraught with difficulties, because a Bayesian model requires the specification of a probability distribution on a given parameter space, and nonparametric models effectively change the dimension of this parameter space in dependence on the sample observation. Ferguson (1973) solved the problem by observing that, since nonparametric models require an a priorily unbounded number of parameters, they can be regarded as parametric models with an infinite-dimensional parameter space. A Bayesian model can be defined by specifying a parameter distribution (a *prior*) on the infinite-dimensional space. Models of this type are now generally referred to as *nonparametric Bayesian models.* The particular model proposed by Ferguson (1973), the *Dirichlet process*, randomly generates infinite-dimensional quantities that are probability distributions on a suitable sample space. It was originally applied to density estimation, and one property of the model that has long worried statisticians is that distributions drawn from it are discrete, even if defined over a continuous sample space. Roughly speaking, even if the underlying data distribution is smooth, the Dirichlet process estimate will always consist of a series of spikes. Much work in nonparametric Bayesian statistics has been devoted to modify the Dirichlet such that it yields smooth estimates (see Walker *et al.*, 1999, for an overview). More recently, researchers in machine learning realized that the Dirichlet process' supposed shortcoming makes it ideally suited for clustering problems: Regarded as a distribution on parameter space, each spike represents the parameter of one cluster. The number and placement of spikes estimated with a Dirichlet process depends on the data, such that the distribution can be used to construct nonparametric clustering models which represent each component by a parametric distribution, but do permit estimation of the number of clusters from data (Ferguson, 1973; Antoniak, 1974).

Sufficient statistics and exponential families. Exponential families are, roughly speaking, probability models completely specified by a statistic of the data that has complexity bounded with respect to sample size (Schervish, 1995). That is, if there is some function of the data with values in a finite-dimensional space, and all information the data contains about the model is summarized by the function value regardless of sample size, the model is an exponential family, and vice versa (Pitman, 1936; Koopman, 1936). Exponential families are a common theme in the machine learning literature, in particular in the context of graphical models. Particular emphasis has been devoted to their geometric and convex-analytic properties, and the ensuing consequences for parameter inference, such as "moment matching" equations (Wainwright and Jordan, 2003, provide an overview). In the work presented here, the emphasis is on sufficient statistics, and exponential families arise, not so much for their information geometric properties, but as the probability models defined by sufficient statistics. The principal importance of sufficient statistics is threefold: First, in the context of Bayesian systems, they give rise to conjugate (hence solvable) models, and define the mapping from prior and data to the posterior in an interpretable and generic fashion. Second, in the context of nonparametric models, definition of a model requires a rule of how additional observations (hence additional degrees of freedom) are to be incorporated into the model. Sufficient statistics define such rules: If the statistic is of first order, a new data point is incorporated without affecting any other data. If it is of second order, such as the classical kernel methods in machine learning, incorporation of a new data point requires computation of pairwise interaction estimates with each individual point previously observed. Third order corresponds to triplet interactions, etc. And finally, we will argue that in infinite-dimensional cases for which a Bayesian posterior is not representable by a Bayes equation, sufficient statistics provide an alternative, explicit representation of the posterior distribution in conjugate models.

Structural constraints. Structural constraints considered here are spatial and temporal regularity assumptions on the input data – prior assumptions that adjacent points in images tend to belong to the same segment, or that the image structure in a video sequence changes smoothly over time. In contrast to “structured output learning” methods, the structural constraints are an input assumption rather than a learning target. The constraints effectively winnow down the size of the solution space, by excluding or discarding as improbable solutions which are not sufficiently smooth. This is exploitable for algorithmic efficiency, as it decreases the size of neighborhoods in solutions space that have to be covered by local search methods.

Stochastic processes. Stochastic process models arise naturally in Bayesian nonparametrics, since nonparametric models require an arbitrarily large number of degrees of freedom. An unbounded number of degrees of freedom translates into infinite-dimensional parameter spaces, and Bayesian methods require the definition of probability measures on such parameter spaces (see e. g. Schervish, 1995). The study of infinite-dimensional probability models is the domain of stochastic process theory, as reflected by the “process prior” terminology of Bayesian nonparametrics. Development of these methods has progressed to a point where they are applicable in a variety of modeling tasks without actually resorting to the mathematical theory of stochastic processes, by drawing on a given set of well-studied available models, in particular Dirichlet, Gaussian and Levy processes (Ghosh and Ramamoorthi, 2002; Rasmussen and Williams, 2006). Since part of this thesis is concerned with the generic construction of nonparametric models, we will have to make more explicit use of some basic notions of the mathematical theory. It is not generally possible to find a closed-form, functional representation of the probability measure defining an infinite-dimensional random process. Stochastic process theory does, however, provide a general way of defining an infinite-dimensional process distribution in terms of an infinite number of finite-dimensional distributions (e. g. Loève, 1977a; Bauer, 1996). If these finite-dimensional distributions can be specified in a common functional form, a representation of the process is obtained which is, in many regards, the closest general analogue to the representation of a probability distribution by a closed-form distribution function. Our key motivation for resorting to this type of representation is that it is ideally adapted to nonparametrics: The finite-dimensional distributions employed in the representations of the process are its finite-dimensional marginals – which are precisely the distributions we actually have to work with in Bayesian nonparametrics when only a finite number of measurements is observed. A central question, addressed in Ch. 5, will be how this technique of representation and marginalization ties in with the concepts of conditioning, sufficiency, and conjugacy, which are central to the application of process models to Bayesian problems.

1.1.2 Contributions

1. **Rank data clustering**

   The term “ranking” refers to a list of items, ordered by a (usually human) subject in order of preference. Each observation of a rank data set is an ordered list of preferences. This kind of data arises for example in psychological experiments or market surveys, and is increasingly collected automatically, for example by web servers monitoring user behavior. By clustering rank data, a data set can be scanned for groups of people with similar preference behavior. In real-world rank data sets, rankings are typically partial, i. e. each person in the survey ranks only some out of the total number of items. Available clustering models for rankings (e. g. Murphy and Martin, 2003) cannot process partial rankings of varying lengths, and the computational cost of model estimation algorithms grows super-exponentially in the total number of ranked items. Common practice in the literature is to discard all rankings in the data except the complete ones. This does, in general, introduce an unnecessary sample bias. The computational cost limits applicability of algorithms to short rankings (such as seven or eight items). In Ch. 3.1, we show how a decomposition of the sufficient statistic of the most widely used rank data model, the Mallows distribution, can be used to construct a clustering model capable of simultaneously representing partial rankings of different lengths. The decomposition also gives rise to a closed-form solution for the model’s partition function, resulting in an estimation algorithm each iteration of which scales linearly in the number of items. The efficiency of the inference algorithm makes the model applicable to rankings of hundreds of items.

2. **Constrained Bayesian nonparametric models.**

   To model smooth segmentations of images, we propose a Dirichlet process with smoothness constraints for spatial data. Smoothness is enforced by a Markov random field. The Dirichlet process generates random values on the nodes of a Markov random field graph, which aggregate into a random number of categories and couple along the graph edges. By increasing the coupling strength, the model is forced into smoother solutions. We show that the conjugate relation between the parametric sampling model (the likelihood component of the DP mixture) and the DP base measure is preserved despite the Markov random field interaction. We also show that the constrained model can be sampled as efficiently as an unconstrained DP mixture, by means of a collapsed Gibbs algorithm.

3. **Nonparametric dynamic linear models.**

   A particularly interesting aspect of DP mixtures is that these models cannot only estimate a model order for a given data set, but can adjust it if the data changes over time. This is the case in the video segmentation problem, where a clustering solution (a segmentation) has to be computed on each consecutive frame. We introduce a dynamic DP model capable of performing such model order adaptation. We derive a multiscale Gibbs sampling algorithm capable of processing the large amounts of data arising in video applications. The model builds on the conjugate relation between the Dirichlet process prior and its posterior (as discussed in more generality in Ch. 5), and is shown to be the nonparametric (infinite-dimensional) analogue of the exponential family dynamic linear models commonly used in Bayesian forecasting (West and Harrison, 1997).

4. **Algorithms for image and video segmentation.**

   Based on the models described above, we develop segmentation algorithms for noisy images and video sequences. We study the application of DP mixture clustering under spatial smoothness constraints to the segmentation of noisy imagery, such as remote sensing radar data. We apply the dynamic DP mixture to the video segmentation problem, and show how inference can be conducted efficiently by multiscale sampling techniques. We also study the application of mixture approximations to the characteristic gamma distributions of synthetic aperture radar data to derive SAR image segmentation algorithms that are efficient and robust with respect to preprocessing of the data. Our nonparametric Bayesian approach to model order selection avoids the notion of the "true" number of segments in an image. Depending on the purpose of the segmentation, the method provides a *level of cluster resolution* as a scalar parameter. Due to the properties of the model, dependence of the number of clusters on the parameter is – for reasonably well-distinguishable segments – not linear, but resembles a step function. For a given image, most choices of the number of clusters do not result in a consistent partition of the image. The model tends to reproduce one consistent solution as the parameter gradually increases, and then to jump directly to the next consistent choice. This effect is additionally pronounced by the application of smoothness constraints, which serve to emphasize the separation of clusters.

5. **Construction principles for nonparametric Bayesian models.**

   The final chapter of the thesis addresses a fundamental theoretical question, how nonparametric Bayesian methods may be generalized beyond the usual Gaussian and Dirichlet process models. Both Gaussian and Dirichlet processes can be regarded as prior distribution on infinite-dimensional parameter spaces. The "nonparametric" character of Bayesian methods based on such priors is that only a finite (but variable) number of degrees of freedom is used to account for a given finite set of observations. Both models are also constructed mathematically in a similar manner, by specifying the properties of the infinite-dimensional model by means of its finite-dimensional marginals. For the Gaussian and Dirichlet processes, these marginals are Gaussian and Dirichlet distributions, respectively. We consider the construction of nonparametric models from arbitrary finite-dimensional marginals. In addition to the construction of the probability measures themselves, we will be interested in what the finite-dimensional marginals may tell us about those properties of the infinite-dimensional Bayesian inference process. These are not properties of individual measures, but of Bayesian equations, and so we will study the specification of infinite-dimensional prior-posterior pairs in terms of finite-dimensional Bayesian equations. Some key properties that we establish are the following:

   - Complete Bayesian equations can be extended to the infinite-dimensional case in essentially the same manner as individual distributions.
   - If (and only if) all finite-dimensional Bayesian equations required to specify the infinite-dimensional model are conjugate, then so is the infinite-dimensional model. This means, roughly speaking, that finite-dimensional models with a posterior of closed analytic form determine an infinite-dimensional posterior of closed analytic form. A consequence is that (under minimal regularity conditions) nonparametric models with analytic posteriors can be constructed only from exponential family models.
   - If each finite-dimensional model has a sufficient statistic, which is once again the case if and only if it is an exponential family model, then the infinite-dimensional model also has a sufficient statistic, and the functional form of the latter one can be derived from its finite-dimensional counterparts. Since the Bayesian inference process in conjugate models, i. e. the mapping that takes the prior parameters and the data to the parameters of the corre-

> sponding posterior, is completely defined by the sufficient statistic, this provides a generic, constructive specification of the nonparametric model posteriors. Moreover, this specification of the inference process remains applicable even for infinite-dimensional models that do not admit an explicit Bayesian equation. In the finite-dimensional case, the Bayesian equation gives an explicit (though not necessarily tractable) formula for how the prior must be modified, given data, to obtain the corresponding posterior. Such an update equation cannot generally be derived for infinite-dimensional model (despite the fact that the posterior exists), which poses an additional hurdle for Bayesian estimation. The infinite-dimensional analogue of the sufficient statistic provides an alternative way of explicitly specifying the posterior.

We illustrate the general construction results by construction examples, including both the reconstruction of familiar standard models by means of the results sketched above, and the construction of a new model.

1.1.3 Organization

Chapter 2: Background

The brief summary of mathematical and algorithmic preliminaries provided here serves both as a survey of existing work, and to provide a common formal framework for the remainder of the thesis. In order to treat nonparametric Bayesian systems as infinite-dimensional parametric models, we use the representation of parametric models as conditional probability measures, and discuss sufficient statistics, exponential families, Bayesian nonparametrics, and latent variable algorithms from this points of view.

Chapter 3: Clustering with Parametric Mixtures

Contributions are presented in the thesis organized according to model properties, rather than applications. Chapter 3 discusses clustering algorithms based on classical mixture models, for applications to rank data (Sec. 3.1) and synthetic aperture radar imagery (Sec. 3.2).

Chapter 4: Clustering with Nonparametric Mixtures

This chapter develops clustering algorithms with model order selection. This includes clustering of spatial data under Markov random field smoothness constraints (Sec. 4.1), with applications to segmentation of noisy image data, and clustering with model order adaptation, with application to video segmentation (Sec. 4.2).

Chapter 5: Construction of Nonparametric Bayesian Models

The final chapter develops theoretical contributions. We derive construction techniques for nonparametric Bayesian models, give criteria for when these models admit a conjugate posterior, and study the existence and properties of sufficient statistics of the models so obtained. A number of construction examples are included to illustrate the method and its scope.

Appendix

The chief purpose of the appendix is to summarize for reference a number of basic results from probability theory, which are drawn upon by proofs in Ch. 5, but do not fit well into the context of Ch. 2. These results are collected in Sec. A. Additionally, we provide a brief review of conditional probability measures and conditional expectations (Sec. B), and of dominated and undominated probability models (Sec. C).

Chapter 2

Background

This chapter provides a summary of background material and literature references. A rich and diverse literature exists on mixture models and clustering, Bayesian models, Bayesian nonparametrics, and the corresponding inference algorithms. Unfortunately, available references cover only part of the material of concern in the following, from varying perspectives, and often at levels of mathematical abstraction that are either too restrictive or too advanced for the problems considered in this work. The presentation in this chapter is an attempt to bring together notions from different parts of the literature, and to emphasize their common properties.

A note on the use of measure theory. The present dissertation is a thesis in machine learning, and parts of the following presentation are rather abstract by the standards of the field. This level of abstraction has a specific purpose, which is to present parametric and nonparametric (or finite- and infinite-dimensional) Bayesian models within a common framework. One of the underlying themes of the following chapters will be that both types of models share a number of important properties, and their application for modeling purposes follows similar rules, a fact clarified by a joint representation. Since some models, such as the Dirichlet process, do not admit a density representation, a joint formulation has to generalize beyond densities, and involves some basic notions of measure theory, in particular probability measures, abstract conditional expectations, and regular conditional probabilities. Wherever the joint representation is not an issue, the more familiar density formalism will be used.

2.1 Notation

The lion share of notation required in the following concerns probability models and random variables. Random variables are defined on a common, abstract probability space, which will always be denoted $(\Lambda, \mathcal{A}, \mathbb{P})$. All random variables are measurable mappings from this common space into their respective sample spaces. Random variables will be written upper-case, and sample spaces and their σ-algebras will be indexed by the associated random variable, for example $X : (\Lambda, \mathcal{A}) \to (\Omega_x, \mathcal{A}_x)$ for a random variable X with sample space Ω_x. The values in the sample space assumed by X are denoted by the corresponding lower-case letter x. Whenever random variables are endowed with a particular meaning, such as observations or parameters, X denotes observations, Θ a parameter variable, and Y a hyperparameter. Arbitrary σ-algebras are denoted $\mathcal{A}, \mathcal{C}$ etc. A symbol $\mathcal{B}$ always denotes a Borel σ-algebra. The probability measure μ of a random variable X is the image $\mu = X(\mathbb{P})$. When dealing with multiple random variables in the same context, measures are indexed by their variable as μ_X or μ_Θ.

Conditional probabilities (cf. App. B) are written $\mu(X|\Theta)$, where X may be substituted by a measurable set and Θ by a σ-algebra, depending on the context. Elements of the abstract probability space Λ are denoted ω, such that the conditional probability is $\mu(X|\Theta)(\omega)$ when regarded as a function. If $\mu(X|\Theta)$ has a conditional density, it is written $p(x|\theta)$. The letter s generally denotes a sufficient statistic, and a capital S the random variable $S := s(X)$. Expectations are denoted $\mathbb{E}$, and conditional expectations (in the abstract Kolmogorov sense, see App. B) in the form $\mathbb{E}[X|\mathcal{C}]$. An expectation may be indexed by the random variable or measure w. r. t. which it is computed, e. g. $\mathbb{E}_X[\,.\,]$ or $\mathbb{E}_{\mu_{\Theta|X}}[\,.\,]$. The index denotes the random variable rather than a parameter: $\mathbb{E}_\Theta[\,.\,]$ denotes an expectation computed by integrating w. r. t. Θ, *not* the expectation of some variable X for parameter value Θ.

2.2 Parametric Models

The familiar form of a parametric probability model is that of a parameterized family of densities $p(x|\theta)$. The prototypical example is the Gaussian family indexed by mean and variance. More generally, when abstracting from the density representation, a parametric model is parameterized family of probability measures.

2.2.1 Parametric Families

A key decision in the definition of a parametric model is whether or not to regard the parameter as a random variable. If the parameter is non-random, it can be understood either as an index without functional meaning, specifying an element within the density, or as parameter of the measure regarded as a function. If it is considered a random variable, it is meaningful to ask how the parameter and the observation variables couple. Regarding the parameter as a random quantity is the most basic characterization of the Bayesian approach to estimation. Without further restrictions, a parametric family may be arbitrarily complex (choose an arbitrary set with each element indexed by itself). In the following, our notion of a parametric model will be Bayesian, in so far as the parameter will be regarded as a random variable, and elements of the model family are indexed by conditioning.

Definition 1 (Parametric family)**.** Let $(\Lambda, \mathcal{A}, \mathbb{P})$ be an abstract probability space, and $(\Omega_x, \mathcal{B}_x)$ and $(\Omega_\theta, \mathcal{B}_\theta)$ two Borel spaces. Let $X : (\Lambda, \mathcal{A}) \rightarrow (\Omega_x, \mathcal{B}_x)$ and $\Theta : (\Lambda, \mathcal{A}) \rightarrow (\Omega_\theta, \mathcal{B}_\theta)$ be two random variables, and $\mu := X(\mathbb{P})$ the image measure of $\mathbb{P}$ under X. Then the conditional distribution $\mu(X|\Theta)$ is called a *parametric family* of models. For any $\theta \in \Omega_\theta$, the measure $\mu(X|\Theta = \theta)$ will be denoted $\mu_{X|\theta}$.

The most convenient case is when the sample spaces are *Polish spaces*. A Polish space is a complete separable metric space. The concept of a Polish space will be discussed in more detail in Sec. 2.4.2. For the moment, it is sufficient to say that almost any set usually encountered as a sample space, including Euclidean spaces, finite and countable discrete spaces, and even (separable) Banach and Hilbert spaces, are all Polish. The advantage of assuming a Polish sample space for a parametric model as defined above is that the model can be guaranteed to be a regular conditional probability (cf App. B), that is, we can assume that the parametric model $\mu(X|\Theta)$ is a probability measure on X for every possible value of Θ. Sample spaces will always be assumed to be Polish if they are finite-dimensional. Infinite-dimensional sample spaces can not generally be assumed as Polish (if the dimension is not countable). In classical, i. e. non-Bayesian statistics, a parametric family is a parameterized family of functions which constitute measures for each possible value of the parameter. The definition above is more restrictive regarding the possible complexity of the model, which as a conditional expectation on a Borel space is equivalent to a Markov kernel.

The most convenient representation for a parametric family of measures is as a family of densities. If such a representation exists for all elements of the family w. r. t. a single, common reference measure, the family is called *dominated* by the reference measure (cf App. C for a precise definition). For example, the Gaussian family can be represented by the familiar bell-curve densities w. r. t. Lebesgue measure, so the family of Gaussians is dominated by Lebesgue measure. Considering undominated models is not usually important for finite-dimensional problems, but becomes relevant in the infinite-dimensional case of Bayesian nonparametrics. Many nonparametric Bayesian models (including the Dirichlet process posterior on the real line) are not dominated, and cannot be properly represented as a density.

Definition 2 (Parametric family of densities)**.** Let $\mu(X|\Theta)$ be a parametric family, such that $\mathcal{B}_\theta$ contains singletons. Assume that $\{\mu(\,.\,|\{\theta\})|\theta \in \Omega_\theta\}$ is dominated by some measure ν on $(\Omega_x, \mathcal{B}_x)$. Then the set $\{p_{X|\theta}|\theta \in \Omega_\theta\}$ of conditional densities

$$p_{X|\theta} = \frac{d\mu(\,.\,|\{\theta\})}{d\nu} \tag{2.2.1}$$

will be called a *parametric family of densities.*

For purposes of statistical estimation, the implication of the term "parametric model" is that the complexity of the model is bounded with respect to sample size. Common examples are models with parameters taking values in some vector space of fixed, finite dimension. Such bounded model complexity guarantees an arbitrary amount of observed information to be available per model degree of freedom in the asymptotic case, as opposed to non-parametric models (such as Parzen window estimation), for which the number of model parameters grows with sample size. The restrictions of parametric models typically lead to faster convergence rates of estimates, both in the Bayesian sense (posterior convergence) and in the classical sense (convergence of point estimators).

2.2.2 Sufficiency

A measurable function of observed data is called a *statistic.* Statistics are chosen in data analysis to filter out a certain property of the data (such as the variance statistic, which provides a simple quantification of the data's scatter). Parameters in parametric models are typically chosen to correspond to some statistic of the data. Information about the data not resolved by the parameter statistic is encoded in the model class. The information captured by the statistic discriminates between individual models within the class. If a statistic captures all information the data may contain about the parameter (and hence about the element of the model class), it is called a *sufficient statistic.* Since Bayesian and classical estimation follow different notions of how a model is to be determined, they suggest different notions of sufficiency. In the classical case, the sufficient statistic has to completely specify a particular element of the model class. In the Bayesian setting, it has to completely specify a posterior distribution over models. The two cases turn out to be equivalent, provided that the model class is dominated.

Basic definition

"A statistic", writes Fisher (1922),

> "...satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated."

Translating Fisher's requirement into a probabilistic definition is a bit subtle, because in the classical view, θ is not a random value. Any probabilistic formulation has to be stated in terms of sampling values of X, instead of the parameter. Suppose the sufficient statistic s is a measurable mapping on Ω_x with values in some space Ω_s. That is, $X : (\Lambda, \mathcal{A}) \rightarrow (\Omega_x, \mathcal{B}_x)$ and $s : (\Omega_x, \mathcal{B}_x) \rightarrow (\Omega_s, \mathcal{B}_s)$. Let S be the random variable $S := s(X)$ (the composition mapping $S = s \circ X$), and suppose that it assumes the specific value $S = s_0$ on a given sample. Fisher considers information about the parameter provided *by a sample*, meaning the definition only has to resolve information about adjustments of θ which would change the distribution of X. If $S = s_0$ provides all such information, knowing S completely specifies the distribution of X. In terms of densities, the Fisher definition is then formalized by requiring

$$p(x|S = s, \theta) = p(x|S = s) \; . \tag{2.2.2}$$

Just for the present context (and nowhere else in this thesis) we have to stress that θ is no random variable, so we write a probability measure with parameter θ as $\mu(X; \theta)$, and as $\mu(X|S; \theta)$ when conditioned on S. The definition looks most familiar for densities, but existence of densities is not essential: Replace the parameterized set $p(\,.\,|\,.\,;\theta)$ of densities in (2.2.2) by a parameterized set of measures $\mu(\,.\,|\,.\,;\theta)$. Let $\mathcal{M} = \{\mu(\,.\,;\theta)|\theta \in \Omega_\theta\}$ be an indexed family of measures. Though the eventual purpose is to consider a parametric family, in which θ represents the value assumed by a parameter variable, θ is for now considered simply as an index that identifies an element of $\mathcal{M}$. Likewise, the index range Ω_θ will at first be assumed to be an arbitrary set. The following definition is due to Halmos and Savage (1949).

Definition 3 (Sufficient statistic)**.** Let $(\Omega_x, \mathcal{B}_x)$ and $(\Omega_s, \mathcal{A}_s)$ be measurable spaces, and $\mathcal{M} = \{\mu(\,.\,;\theta)|\theta \in \Omega_\theta\}$ a family of probability measures on

$(\Omega_x, \mathcal{B}_x)$. Let $S : (\Omega_x, \mathcal{B}_x) \to (\Omega_s, \mathcal{A}_s)$ be a measurable map. Then S is called a *sufficient statistic* for $\mathcal{M}$ if there is a Markov kernel $k : \mathcal{B}_x \times \Omega_s \to \mathbb{R}_{\geq 0}$ such that for all θ and all $B \in \mathcal{B}_x, s \in \Omega_s$,

$$\mu(B|S = s; \theta) = k(B, s) \qquad \mu(\,.\,;\theta)\text{-a.e.} \tag{2.2.3}$$

The definition states that the conditional probability given S does not depend on which member of the family is considered: There is a single conditional probability function (the Markov kernel k) that simultaneously describes conditioning on S for all random variables X_θ. It matches Fisher's intuitive definition, as quoted above, by reading equation (2.2.3) from right to left: If the value of s is known, we can sample observations of X from $k(B, s)$, and knowing θ in addition, or changing the value of θ on the left-hand side of the equation, does not affect the conditional distribution. Since conditioning on S means conditioning on the σ-algebra $\sigma(S) = S^{-1}(\mathcal{B}_s)$, the definition immediately generalizes from sufficient statistics to sufficient σ-algebras (by substituting an arbitrary sub-σ-algebra $\mathcal{C} \subset \mathcal{B}_x$ for $\sigma(S)$).

Definition 4 (Sufficient σ-algebra)**.** Let again $\mathcal{M} = \{\mu(\,.\,;\theta)|\theta \in \Omega_\theta\}$ be a set of probability measures on a Borel space $(\Omega_x, \mathcal{B}_x)$. A σ-algebra $\mathcal{C} \subset \mathcal{B}_x$ is called *sufficient* for $\mathcal{M}$ if there is a Markov kernel $k : \mathcal{B}_x \times \Omega_x \to \mathbb{R}_{\geq 0}$ such that, for all $B \in \mathcal{B}_x$,

$$\mu(B|\mathcal{C}; \theta)(x) = k(B, x) \qquad \mu(\,.\,;\theta)\text{-a.e.} \tag{2.2.4}$$

This abstracts quite a bit from the idea of a statistic of data being sufficient for the estimate of a parameter: A system of sets now is sufficient for a set of measures. The intuitive meaning of the definition is that the level of resolution of the σ-algebra $\mathcal{B}_x$, which describes observations, can be coarsened to that of the smaller σ-algebra $\mathcal{C}$, without losing any information relevant for the discrimination between different measures in $\mathcal{M}$. Consequently, whenever a given $\mathcal{C}$ is sufficient for $\mathcal{M}$, any σ-algebra $\mathcal{C}'$ of finer resolution ($\mathcal{C} \subset \mathcal{C}'$) should be sufficient as well. This is indeed true if the set $\mathcal{M}$ is dominated, but as we will discuss below, not in the undominated case.

For dominated families, sufficiency can be characterized by the Neyman factorization criterion, which has become so well-known that it is used as a definition of sufficiency in many texts. The theorem is reproduced here only for the sake of completeness, but will not actually play an explicit role in our applications of sufficiency.

Theorem 5 (Neyman factorization criterion)**.** *Let $\{p_{X|\Theta}(\,.\,|\theta)|\theta \in \Omega_\theta\}$ be a parametric family of densities. Then a statistic S is sufficient for Θ if and* *only if there are functions* g_1, g_2 *such that*

$$\forall \theta \in \Omega_\theta : p_{X|\Theta}(x|\theta) = g_1(x) g_2(S(x), \theta) \ . \qquad (2.2.5)$$

Sufficiency in Bayesian models

The above definition of sufficiency can be regarded as classical rather than Bayesian, not only for eventually being due to R. A. Fisher, but also because it implies the notion of a particular value assumed by the parameter. The Bayesian way of stating that knowledge of $S(X)$ conveys all information about the parameter is to say that the parameter's posterior given X is specified completely by $S(X)$, as in the following definition. To emphasize independence of the sufficient statistic from the choice of the prior distribution, the parametric family, usually written as $\mu(X|\Theta)$, is written as a conditional probability given a σ-algebra.

Definition 6 (Bayesian sufficiency)**.** Let $\mu(X|\mathcal{C})$ be a parametric family as in Def. 1, where $\mathcal{C}$ is a sub-σ-algebra or $\mathcal{A}$, and let $S : (\Omega_x, \mathcal{B}_x) \to (\Omega_s, \mathcal{A}_s)$ be a measurable map. Then S is called a *sufficient statistic* for the parametric family if, for any parameter variable $\Theta : (\Lambda, \mathcal{A}) \to (\Omega_\theta, \mathcal{B}_\theta)$ satisfying $\sigma(\Theta) = \mathcal{C}$,

$$\mu(\Theta|X) = \mu(\Theta|\sigma(X \circ S)) \qquad \Theta(\mathbb{P})\text{-a.e.} \qquad (2.2.6)$$

This definition basically states that S completely specifies the posterior for the chosen family, regardless of the choice of the prior. The distinction in notation between $\mu(X|\mathcal{C})$ and the customary $\mu(X|\Theta)$ is made because the choice of the prior is limited by the choice of the parametric family: The prior is defined by Θ, as the image measure $\mu_\theta := \Theta(\mathbb{P})$. The parametric family is defined by conditioning on the σ-algebra $\mathcal{C}$, *not* on the measure μ_θ. (In terms of density models, the mathematical form of a model $p(x|\theta)$ depends only on the values assumed by θ, not on how probable these values are to occur.) We therefore have to define the parametric family first, by choice of $\mathcal{C}$, and then limit our choice of priors to those induced by random variables which generate just this $\mathcal{C}$. This is a subtlety invisible in common density models, where $\mathcal{C}$ is such that it contains all singletons, and the conditional model is chosen by conditioning the density pointwise.

The two definitions of sufficiency (classical and Bayesian) are equivalent if the parametric family is dominated, but can differ in the undominated case, which is relevant in particular for Bayesian nonparametrics. The following theorem is given in this form by Schervish (1995), but is essentially due to Blackwell and Ramamoorthi (1982).

Theorem 7 (Classical and Bayesian sufficiency)**.** *Classical sufficiency (Def. 3) implies Bayesian sufficiency (Def. 6), with equivalence if there is a σ-finite measure ν on $(\Omega_x, \mathcal{B}_x)$ such that $\mu_{X|\theta} \ll \nu$ for all $\theta \in \Omega_\theta$.*

If the family $\mathcal{M}$ of parametric models is not dominated, two key properties are no longer guaranteed: Bayesian sufficiency does not imply classical sufficiency, and super-σ-algebra of a sufficient σ-algebra need not be sufficient. The implication of a statistic being sufficient in the Bayesian, but not classical sense is that the Bayesian model is unable to resolve at least some cases which are distinguished as different by the classical model. Blackwell and Ramamoorthi (1982) give a hypothesis testing example for which the two notions of sufficiency differ – with the consequence that there is no classical test achieving zero error probability, whereas the Bayesian version of the test always results in a Bayes-sufficient result with probability 1. The Bayesian is always certain to be right, the classical test is always uncertain.

Remark 8 (Sufficiency as a compression property)**.** In general, a sufficient statistic for a given model is not a single function s, but actually a collection of functions s_n, one for each sample size $n \in \mathbb{N}$. For an arbitrary parametric model, we have to expect the range of s_n to get more and more complex as the sample size n increases. References to "the" sufficient statistic implicitly assume that there is a function $s : \Omega_x \to \Omega_s$, where Ω_x is the sample space and Ω_s some space of finite dimension, such that s_n is computable as the arithmetic average

$$s_n(x_1, \ldots, x_n) := \frac{1}{n} \sum_{i=1}^{n} s(x_i) \ . \tag{2.2.7}$$

The implication is that the range of s_n is always contained in Ω_s, regardless of the sample size. Provided that the dimension of Ω_s is finite, this is a *compression property*: A sample $x_1, \ldots, x_n$ has $n \cdot \dim(\Omega_x)$ degrees of freedom, but the information extracted by the sufficient statistic can always be summarized in a vector of $\dim(\Omega_s)$ dimensions.

From an abstract point of view, every parametric model admits a sufficient statistic: For any sample $x_1, \ldots, x_n$, choose s_n as the identity mapping on sample space. The statistic "extracted" from the sample is then the sample itself, and trivially preserves all information contained in the observation. Apparently, in this case, there is no compression, and the dimension of the statistic's range grows with sample size. When we refer to sufficient statistics in the following, we will always assume s_n to take the form of an arithmetic average. At first glance, this may seem a severe restriction, but actually constitutes almost no loss of generality: Roughly speaking, a model that admits a sufficient statistic of bounded complexity (dimension of Ω_s bounded w. r. t. sample size) also admits a sufficient statistics representable as an average. This is a consequence of the so-called Pitman-Koopman lemma, to be discussed in Sec. 2.2.4. Parametric models relying explicitly on the average representation of a sufficient statistic are called exponential family models, which will be defined in the next section.

2.2.3 Exponential Families

A generic approach to the construction of probability models is to decide which properties of a data source are of interest, and to define a set of statistics that measure these properties. The model is then defined such that it resolves those properties measured by the statistics, and only those – which implies that the chosen statistics are sufficient for the model. The parametric models constructed in this manner are the *exponential family models.* Given a set of statistics, there is a generically defined class of exponential families for which these statistics are sufficient. Conversely, if a model has sufficient statistics (the complexity of which does not grow with sample size), then it is an exponential family model. Another way of saying that a model resolves only information contained in the sufficient statistics is to say that, given the constraint that the statistics assume certain values, the model is maximally indetermined (or maximally random). Maximal randomness can be formalized as maximal entropy, and exponential families are maximum entropy models, with constraints specified by the sufficient statistics. In statistical physics, such models are known as Gibbs distributions. Much and more has been written about exponential families, their geometric properties and their application in Bayesian estimation. Key references include Barndorff-Nielsen (1970, 1973, 1978); Efron (1978); Brown (1986); Diaconis and Ylvisaker (1979). In the following, we will regard exponential families largely as a by-product of sufficient statistics, and largely neglect their geometric and convex analytic interpretation.

Consider data $x_1, \dots, x_n$ generated conditionally independent given the value θ of some random quantity Θ. Assuming that the conditional distribution $\mu(X|\Theta)$ is dominated, the joint conditional distribution of the observations x_i has conditional density $f(x_1, \dots, x_n|\theta)$ with respect to a carrier measure ν. By conditional independence, it can be represented as

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f_{X|\theta}(x_i|\theta) \ . \tag{2.2.8}$$

Any conditional density can be rewritten as $f_{X|\theta}(x|\theta) = \frac{1}{Z(\theta)} \exp(-H(x|\theta))$, also known as an *energy representation* with energy function H, such that for the joint density,

$$f(x_1,\dots,x_n|\theta) = \frac{1}{Z(\theta)^n}\prod_{i=1}^{n}\exp(-H(x_i|\theta))\ . \tag{2.2.9}$$

If a sufficient statistic s is applied to each sample, the component densities can be rewritten, by suitable modification of H, as $1/Z(\theta)\exp(-H(s(x_i)|\theta))$. The simplest possible form of H that correlates $s(x)$ to θ is a bilinear energy $-H(s(x)|\theta) = \langle s(x)|\theta\rangle$, for which the joint density above can be rewritten as

$$f(x_1,\dots,x_n|\theta) = \frac{1}{Z(\theta)^n}\exp\Big(\sum_{i=1}^{n}\langle s(x_i)|\theta\rangle\Big)\ . \tag{2.2.10}$$

The corresponding distribution model can in fact be derived directly from the sufficient statistic, by entropy maximization. (The following derivation is heuristic in so far as it omits relevant smoothness assumptions on the densities.) Let $s:\Omega_x\to\Omega_\theta$ be some statistic, and assume that Ω_θ has an inner product. For a probability density p, denote by $\mathcal{S}$ the entropy functional

$$\mathcal{S}[p] := -\int_{\Omega} p(x)\log(p(x))d\nu(x)\ . \tag{2.2.11}$$

Then the variational problem

$$\max \quad \mathcal{S}[p] \tag{2.2.12}$$

$$\text{s.t.} \quad \mathbb{E}_p\left[s(X)\right] = \theta \tag{2.2.13}$$

is solved by a density of the form

$$p(x) = \frac{\exp(\langle s(x)|\theta\rangle)}{\int\exp(\langle s(x)|\theta\rangle)\nu(dx)}\ . \tag{2.2.14}$$

This is apparently just the bilinear energy case described above. The exponential family for the sufficient statistic s will be defined as follows.

Definition 9 (Exponential Family Model)**.** Let $f_{X|\theta}$ be a parametric family of densities for which Ω_θ has an inner product $\langle\,.\,|\,.\,\rangle$. The family is called an *exponential family* if the conditional density is expressible in the form

$$f_{X|\theta}(x|\theta) = \frac{1}{Z(\theta)}h(x)\exp\left(\langle s(x)|\theta\rangle\right)\ , \tag{2.2.15}$$

where $s:\Omega_x\to\Omega_\theta$ is measurable, $h:\Omega_x\to\mathbb{R}_+$, and $\theta\in\mathcal{T}\subset\Omega_\theta$ such that $\mathcal{T}$ is open and convex.

The definition differs from the maximum entropy model above in the function h. Since h is finite and positive, it can always be absorbed into the carrier measure, resulting in the maximum entropy model for the modified carrier $h(x)\nu(dx)$. An exponential family model in the above sense is therefore uniquely determined be choosing the data domain, the parameter space, the carrier measure, and the sufficient statistic. The integral form $Z(\theta) = \int h(x) \exp(\langle s(x)|\theta\rangle) d\nu(x)$ of Z is a necessary consequence. A more general definition is obtained by substituting the image $\tau(\theta)$ of the parameter θ under some suitable mapping τ. The resulting class of distributions is mathematically equivalent, since τ may be transformed out to obtain an equivalent model on Ω_θ. Some well-known models, such as the Weibull distribution, constitute exponential family models with respect to all parameters only if a parameter transform is allowed – which is, however, just a way of saying that the customary representation of the Weibull involves a parameter transform, which could be eliminated without changing the model's sampling properties. Popular exponential family models include the normal distribution, the gamma and its special cases (χ^2, Erlang and exponential), the binomial, Bernoulli, beta, Poisson, hypergeometric, inverse normal, inverse gamma, negative binomial and Rayleigh distributions (all one-dimensional). Multidimensional examples include Gaussian, Wishart, multinomial, Dirichlet and Mallows models. Bernardo and Smith (1994) and Brown (1986) provide a (partial) taxonomy.

The energy representation (2.2.9) of the model can be interpreted as follows. Assume that the set of all admissible densities is defined by models of the form (2.2.9) such that, for any θ, the energy $H(\,.\,|\theta)$ is in $L_2(\Omega_x)$. The component functions s_i of the vector $s(x) = (s_1(x), \ldots, s_d(x))$ span a linear subspace $\text{span}\{s_1, \ldots, s_d\}$ of $L_2(\Omega_x)$. The exponential family models for s are just those densities whose energies are elements of the finite-dimensional, linear subspace.

2.2.4 Pitman-Koopman Theory

An exponential family always admits a sufficient statistic. The theorem below states that the converse is also true: Parametric models admitting sufficient statistics of fixed dimension are exactly those representable as exponential family distributions. A substantial amount of literature has been devoted to this subject, known as *Pitman-Koopman theory* (or by any combination of the names Pitman, Koopman, Darmois and Fisher). The following version of the Pitman-Koopman result is due to Jeffreys (1961).

Theorem 10 (Pitman-Koopman lemma). *Let the random quantities X_1, $X_2, \ldots$ be conditionally i.i.d. given the value of some random quantity Θ,* *and assume that the conditional distribution $\mu(X_i|\Theta)$ is dominated by a measure ν. Let $f_{X|\theta}$ be the corresponding conditional density. Assume further that the support of $f_{X|\theta}$ is independent of the value of θ:*

$$\forall\theta_1,\theta_2 \in \Omega_\theta : \qquad \textit{supp}\ f(\,.\,|\theta_1) = \textit{supp}\ f(\,.\,|\theta_2) \qquad \nu\text{-}a.e. \quad (2.2.16)$$

Then if there is a sufficient statistic $s : \Omega_x \to \Omega_s$ for the model and if Ω_s has finite dimension, $f_{X|\theta}$ is an exponential family model.

The above statement of the lemma implicitly assumes observations to be exchangeable, by conditional independence. The preconditions of the theorem can be modified in a variety of manners. Some proofs do not rely on a θ-invariant support of the density, whereas others omit the assumption of identically distributed observations. Versions omitting the above conditions substitute others, often involving analytic smoothness properties of the densities. The one requirement common to all flavors of the result is the fixed, finite dimension of Ω_s, and hence the constant complexity of the sufficient statistic w. r. t. sample size.

2.2.5 References

The concept of sufficiency is due to Fisher (1922), and thus predates Kolmogorov's work on abstract conditional expectations (Kolmogorov, 1933). It received a rigorous treatment in the light of conditional distributions by Halmos and Savage (1949), for the dominated case. The Neyman factorization theorem is in part due to Fisher (1922, 1934). The name "Neyman theorem" originated with the article by Halmos and Savage (1949), who cite as their source a somewhat obscure work by Neyman (1935), published in Italian in the *Giornale Dell'Istituto Italiano degli Attuari*. Quite probably, later citations of this work are due to the reference in Halmos and Savage (1949), rather than to the original publication. The work of Halmos and Savage (1949) is extended by Bahadur (1954), who devotes a detailed discussion to the problem of domination and is the first to raise the question of whether a refinement of a sufficient σ-algebra (i. e. a σ-algebra containing a sufficient σ-algebra) can be non-sufficient – a counter-intuitive situation, and impossible in the dominated case. The question is answered in the affirmative by Burkholder (1961). An analogue of the Neyman factorization criterion for undominated sufficiency is obtained by Ghosh *et al.* (1981). The problem of minimality in the undominated case was first studied by Basu and Ghosh (1969).

The Pitman-Koopman property was first hinted at by Fisher (1934). Three independent versions of the actual result appeared in the following two years, first by Darmois (1935) in the *Comptes Rendus*, followed by the works of Koopman (1936) and Pitman (1936) in the anglophone literature. Already, the assumptions of the proofs differ. Pitman does not require the densities' support to be parameter-independent, but considers only one-dimensional parameter space. Koopman's result applies to multiple dimensions, but requires parameter-independent densities. Only Koopman explicitly mentions the requirement, tacitly implied by both other authors, that the densities be analytic functions. Numerous modifications have since been obtained, under different conditions on the densities involved. For example, Barankin and Maitra (1963) generalize beyond the case of identically distributed samples. Andersen (1970) considers discrete sample spaces. Other references include Dynkin (1961); Brown (1964); Denny (1964, 1967, 1972); Barndorff-Nielsen and Pedersen (1968); Hipp (1974); Huzurbazar (1976); Lauritzen (1988). A similar result, again under regularity conditions, is available for stochastic processes (Küchler, 1982a,b).

2.3 Bayesian Inference

Bayesian inference treats the model, or the parameter specifying the model, as a random quantity. Since the parameter is random, it has a distribution, and the objective of Bayesian inference is to determine the distribution of the parameter given the data. Application to a given problem requires a suitable representation of the conditional distributions involved, usually in terms of a density (though densities may not be applicable in the Bayesian nonparametric case).

A random model can be specified by choosing a parametric family $\{\mu_{X|\theta}| \theta \in \Omega_\theta\}$ of models, and treating the parameter as a random quantity, as is actually done in Def. 1 of parametric families. Bayesian inference given an observation X then requires determination of the conditional probability $\mu_{\Theta|X}$. The image measure $\mu_\Theta := \Theta(\mathbb{P})$ is called the *prior*, and the conditional probability $\mu_{\Theta|X}$ or $\mu_{\Theta|X_1,\dots,X_n}$ the *posterior*. The parametric model is referred to as the *likelihood* or *sampling distribution*.

2.3.1 The Bayes Equation for Dominated Models

Computation of the posterior is in general a nontrivial task. Like many other problems, it is greatly simplified if the parametric family is suitably dominated: Determining the posterior, as a conditional distribution, becomes much easier if it possesses a density conditional on the observations. That in turn requires the family of all posteriors $\{\mu_{\Theta|X}(\Theta|X = x)|x \in \Omega_x\}$

to be dominated. To actually compute a conditional density, a particular dominating measure has to be chosen. The natural choice is the prior: Since the posterior is a distribution on the parameter, it should preferably be represented w. r. t. the measure on the parameter variable, and that is the prior μ_Θ. The conditional density of the posterior can then be read as a reweighting which the distribution of the parameters undergoes due to observations. To ensure existence of the conditional with respect to the prior, the posterior family must be shown to be dominated by the prior. The following theorem states that, if the parametric model is dominated, then (i) the posterior family is automatically dominated by the prior, and (ii) the density of the posterior with respect to the prior has a generally applicable representation in terms of the parametric family's density. Note that the assumption of the model constituting a parametric family of densities in the statement of the theorem implies that the model is dominated.

Theorem 11 (Bayes equation)**.** *Let $\{p_{X|\theta}|\theta \in \Omega_\theta\}$ be a parametric family of densities, and μ_Θ the prior measure on Θ. Let $N_x \subset \Omega_x$ be the set of all x for which $\int p(x|\theta)\mu_\Theta(d\theta) \in \{0, +\infty\}$. Then for any $x \notin N_x$, the posterior $\mu_{\Theta|X=x}$ has Radon-Nikodym derivative*

$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta|x) = \frac{p_{X|\theta}(x|\theta)}{\int p(x|\theta)d\mu_\Theta(\theta)} \,. \tag{2.3.1}$$

Though the set N_x is a potential liability, it does not actually cause problems, because observations in N_x do not occur:

Lemma 12. *The prior predictive probability of N_x, i. e. the probability of observing $x \in N_x$ under a two-stage sampling model $\theta \sim \mu_\Theta$ and $x \sim \mu(X|\Theta = \theta)$, is zero.*

In Bayesian nonparametrics, the domination assumption often fails. For example, for the Dirichlet process on the real line, the family of posteriors is not dominated by the prior, and cannot be represented by a Radon-Nikodym derivative. Such models require alternative ways of representing the posterior measure, as will be discussed in detail in Chapter 5.

2.3.2 Conjugate Bayesian Models

The posterior density $\frac{d\mu_{\Theta|X}}{d\mu_\Theta}$ in the Bayes equation (2.3.1) exists whenever the parametric sampling model is dominated, but it does not in general possess a closed-form solution, and need by no means be feasible to evaluate for a given problem.

The only known general class of models for which such a closed-form solution exists are exponential family models in combination with their so-called *conjugate priors*. A family of priors is called conjugate to a given sampling model if any posterior under observation is an element of the prior family. This property is commonly referred to in the literature as *closure under sampling* of the prior family. The following definition states precisely the same in formal terms, and looks more complicated then it actually is because it involves a conditional probability (the posterior). As anything containing a conditional, it comes with an "almost surely" caveat, and has to specify to which measure the caveat relates.

Definition 13 (Conjugate prior family)**.** Let $\mathcal{M} = \{\mu_{X|\theta} | \theta \in \Omega_\theta\}$ be a parametric family of models, and $\mathcal{N}$ a family of priors. Then $\mathcal{M}$ and $\mathcal{N}$ are called *conjugate families* if for any $\nu \in \mathcal{N}$ and any observation $x \in \Omega_x$, the posterior $\mu_{\Theta|X}(\Theta|X = x)$ is an element of $\mathcal{N}$, almost surely with respect to $\mu_{\nu,X} = \int \mu_{X|\theta}(X|\theta) d\nu(\theta)$.

Further requirements are necessary to make this property non-trivial, since apparently the set of all probability measures is conjugate to any model family. The term conjugate prior as used in the Bayesian literature typically implies that the prior family is a parametric family, and that there is some well-defined rule that computes the parameter of the posterior from the parameter of the prior and the observed data.

Definition 14 (Conjugate parametric model)**.** Let $\mathcal{M} = \{\mu_{X|\theta} | \theta \in \Omega_\theta\}$. A parametric family $\mathcal{N} = \{\mu_{\Theta|y} | y \in \Omega_y\}$ is a *conjugate parametric family* of priors for $\mathcal{M}$ if, for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \Omega_x$ and $y \in \Omega_y$, there is a $y' \in \Omega_y$ such that $\mu_{\Theta|Y}(\Theta|Y = y')$ is a version of the posterior $\mu_{\Theta|X^n,Y}(\Theta|X_1 = x_1, \ldots, X_n = x_n, Y = y)$.

Sufficiency by Conjugacy

The definition of a conjugate parametric model above is closely related to sufficiency in the Bayesian sense: For each $n \in \mathbb{N}$, define a mapping s_n as follows: For any $x_1, \ldots, x_n \in \Omega_x$ and $y \in \Omega_y$, let y' be a value specifying a posterior within $\mathcal{N}$, and define s_n by $s_n(x_1, \ldots, x_n, y) =: y'$. The posterior is then completely determined by the value of s_n, much like in the definition of Bayesian sufficiency. The difference between the two definitions is that sufficiency requires the posterior to be determined by s under *any* prior. The conjugate case only guarantees a functionally determined posterior for those priors which are in $\mathcal{N}$. The immediate question is whether conjugacy implies sufficiency, or whether it is possible to find a conjugate prior even if the model does not admit a sufficient statistic.

The following lemma states that, for dominated families, conjugacy implies sufficiency if the prior densities in $\mathcal{N}$ do not vanish anywhere on parameter space.

Lemma 15 (Non-degenerate conjugate models admit sufficient statistics)**.** *Let $\mu_{X|\theta}(X|\Theta)$ and $\mu_{\Theta|y}(\Theta|Y)$ be two dominated parametric families on Borel spaces $(\Omega_x, \mathcal{B}_x)$ and $(\Omega_\theta, \mathcal{B}_\theta)$. Let the families be conjugate, in the sense that there is a function $s : \Omega_x \times \Omega_y \to \Omega_y$ such that the posterior for any prior $\mu_{\Theta|y}(\Theta|Y = y)$ in the second family under observation $X = x$ is $\mu_{\Theta|y}(\Theta|Y = s(x, y))$. Then if the density of the prior $\mu_{\Theta|y}(\Theta|Y)$ is strictly positive on $\Omega_\theta \times \Omega_y$, the function $s(\,.\,) := s(\,.\,, y)$ is a sufficient statistic for $\mu_X|\theta(X|\Theta)$ in the classical sense.*

The condition that all prior densities be positive essentially requires the priors to distribute their probability mass over the same region in Ω_θ, since regions in which all prior densities are zero could be removed from the parameter space. This is a similar requirement on parameter space to that imposed by Th. 10 on Ω_x. For lack of a reference on this result, La. 15 is proven below.[1] It is possible to construct pathological examples of conjugate models that do not yield a sufficient statistic: Consider for instance the a set of priors consisting of all Dirac measures on the parameter space Ω_θ, for some smooth sampling distribution. Then for every Dirac concentrated at some $\theta \in \Omega_\theta$, the posterior is again the same Dirac measure, whatever the observations. The class is therefore conjugate. Since the Dirac measures are parameterized by their position, the identity mapping $\mathrm{Id}_{\Omega_\theta}$ completely determines the posterior parameter. Apparently, this does not imply that $\mathrm{Id}_{\Omega_\theta}$ (which does not even depend on the data) is a sufficient statistic for the sampling distribution. However, such examples only exist in cases where the prior is in some way degenerate.

Proof (La. 15). The proof proceeds in two steps. By Def. 3, classical sufficiency requires the existence of a Markov kernel k that satisfies Eq. (2.2.3). Step 1 derives a Markov kernel to serve as a candidate for k. Step 2 then shows that the kernel indeed satisfies Eq. (2.2.3).
Step 1. Let ν_X and ν_Θ be dominating measures for the two families. Define

[1] Though I am convinced that a result comparable to La. 15 should exist somewhere in the literature, I am not aware of a reference. The proof given here is adapted from a proof given by Schervish (1995) for the equivalence of classical and Bayesian sufficiency (cf Th. 7). The measure ρ used in the proof is constructed just as in Schervish's proof. All we have to do is express the density of ρ as a function of the conjugate posteriors. The remainder of the proof, though somewhat lengthy, then follows by elementary computations.

the respective conditional densities as

$$f_{X|\theta} := \frac{d\mu_{X|\theta}}{d\nu_X} \qquad \text{and} \qquad g_{\Theta|y} := \frac{\mu_{\Theta|y}}{d\nu_\Theta} \; . \tag{2.3.2}$$

Because the family $\mu_{X|\theta}$ is dominated, there exists (according to Cor. 59) a measure ρ on $(\Omega_x, \mathcal{B}_x)$ such that (i) $\mu_{X|\theta} \ll \rho \ll \nu_X$ and (ii) ρ has a representation as a countable convex combination of measures in the family. That is,

$$\rho = \sum_{i=1}^{\infty} c_i \mu_{X|\theta_i} \tag{2.3.3}$$

for some countable sequences $\{\theta_i\}_{i\in\mathbb{N}}$ of parameters in Ω_θ and $\{c_i\}_{i\in\mathbb{N}}$ of mixture weights, where $\sum_{i\in\mathbb{N}} c_i = 1$. By the chain rule for Radon-Nikodym derivatives, the density of $\mu_{X|\theta}$ (for any θ) with respect to ρ is

$$\frac{d\mu_{X|\theta}}{d\rho} = \frac{f_{X|\theta}}{\sum_{i\in\mathbb{N}} c_i f_{X|\theta_i}} \; . \tag{2.3.4}$$

The key to the proof is now to express the density $\frac{d\mu_{X|\theta}}{d\rho}$ as a function of the conjugate posteriors, and hence as a function that depends on $s(x,y)$, but not directly on x. Since both families are dominated, the Bayes equation (Th. 11) applies, and the density of the posterior is

$$\frac{d\mu_{\Theta|x,y}}{d\nu_\Theta}(\theta|x) = \frac{f_{X|\theta}(x|\theta) g(\theta|y)}{\int_{\Omega_\theta} f_{X|\theta}(x|\theta) g(\theta|y) d\theta} \; . \tag{2.3.5}$$

To simplify notation, we will write $f(x|y)$ for the integral in the denominator, where we set $f(x|y) := 1$ in cases where the integral takes infinite or zero value. Since the model is conjugate, the posterior density can be expressed by a density in the prior family, as

$$\frac{d\mu_{\Theta|x,y}}{d\nu_\Theta}(\theta|x) = g(\theta|s(x,y)) \; . \tag{2.3.6}$$

The regularity assumption on g (that g does not vanish anywhere on Ω_θ) in order to guarantee that the quotient (2.3.4) can be expressed in terms of the posterior and the prior: Since g is non-zero everywhere the Bayes equation (2.3.5) can be solved for $f_{X|\theta}$, and substitution into (2.3.4) gives

$$\frac{d\mu_{X|\theta}}{d\rho} = \frac{g(\theta|s(x,y))\frac{f(x|y)}{g(\theta|y)}}{\sum_i c_i g(\theta_i|s(x,y))\frac{f(x|y)}{g(\theta_i|y)}} = \frac{g(\theta|s(x,y))}{g(\theta|y)\sum_i c_i \frac{g(\theta_i|s(x,y))}{g(\theta_i|y)}} =: h_y(\theta, s(x,y)) \; . \tag{2.3.7}$$

Note that the function h on the right-hand side is indexed by the hyperparameter y. Since the left-hand side is independent of y, but the y does occur in the argument $s(x, y)$ of h, the form of h must in general be adjusted whenever y changes, to make the overall expression independent of the hyperparameter. To show that $s(\,.\,, y)$ is indeed sufficient for any fixed value of y, we have to show that the Markov kernel k in Eq. (2.2.3) exists. To this end, we fix y and henceforth avoid explicit use of y in all equations, writing $s(x)$ for $s(x, y)$ and h for h_y to avoid awkward notation. Define a new random variable $U := s(X)$ and take the conditional probability of the *image* measure $s(\rho)$, given $U = u$:

$$k(A, u) := \mathbb{E}\left[\mathbb{I}_A | U = u\right] \qquad \text{for } A \in \mathcal{B}_x \ . \tag{2.3.8}$$

Since $(\Omega_x, \mathcal{B}_x)$ is a Borel space, $k(A, u)$ can be chosen as a version that is a Markov kernel. From here on, the proof is largely a matter of computation. *Step 2.* What remains to be shown is that k indeed satisfies Eq. 2.2.3, i. e. that $k(A, u)$ is a version of the conditional probability $\mu_X(A|\Theta = \theta, U = u)$ for all $\theta \in \Omega_\theta$. By definition, this is the case if k integrates as the conditional probability would over all sets C,

$$\int_C k(A, u) ds(\mu_{X|\theta})(u) = \int_C \mu_X(A|\Theta = \theta, U = u) ds(\mu_{X|\theta})(u) \ . \tag{2.3.9}$$

This equality is what we have to show. Note the catch: The integral here is evaluated with respect to the image measure $s(\mu_{X|\theta})$, but the definition of k is with respect to $s(\rho)$. Consider first the expression on the right, which by definition of conditional probabilities is

$$\begin{aligned} &\int_C \mu_X(A|\Theta = \theta, U = u) ds(\mu_{X|\theta})(u) \\ &\qquad \overset{\text{pull-back}}{=} \int_{s^{-1}(C)} \mu_X(A|\Theta = \theta, U = s(x)) d\mu_{X|\theta}(x) \\ &\qquad \overset{\text{cond. prob.}}{=} \mu_{X|\theta}(A \cap s^{-1}(C)) \ . \end{aligned} \tag{2.3.10}$$

The last expression can be rewritten as

$$\mu_{X|\theta}(A \cap s^{-1}(C)) = \int_{s^{-1}(C)} \mathbb{I}_A(x) d\mu_{X|\theta}(x) \ . \tag{2.3.11}$$

Since $s^{-1}(C) \in \sigma(s)$, the variable $\mathbb{I}_A$ integrates over the set $s^{-1}(C)$ just as its conditional expectation $\mathbb{E}\left[\mathbb{I}_A|\sigma(s)\right]$, and we write

$$\int_{s^{-1}(C)} \mathbb{I}_A(x) d\mu_{X|\theta}(x) = \int_{s^{-1}(C)} \mathbb{E}\left[\mathbb{I}_A | U = s(x)\right] d\mu_{X|\theta}(x) \ . \tag{2.3.12}$$

In comparison to the first integral in Eq. (2.3.10), the condition $\Theta = \theta$ has been eliminated from the conditional expectation in the integrand. To substitute k in the integrand, the integral measure has to be transformed to $s(\rho)$. By (2.3.4) and (2.3.7), $d\mu_{X|\theta} = h(\theta, s(x))d\rho(x)$, and by the chain rule for densities,

$$ds(\mu_{X|\theta})(u) = h(\theta, u)ds(\rho)(u) \ . \tag{2.3.13}$$

Rewrite the integral as

$$\begin{aligned}\int_{s^{-1}(C)} \mathbb{E}\left[\mathbb{I}_A | U = s(x)\right] d\mu_{X|\theta}(x) &= \int_C \mathbb{E}\left[\mathbb{I}_A | U = u\right] h(\theta, u)ds(\rho)(u) \\ &\overset{2.3.8}{=} \int_C k(A, u)h(\theta, u)ds(\rho)(u) \ .\end{aligned} \tag{2.3.14}$$

Finally, the integral of k is transformed back to measure $s(\mu_{X|\theta})$, as

$$\int_C k(A, u)h(\theta, u)ds(\rho)(u) = \int_C k(A, u)ds(\mu_{X|\theta})(u) \ , \tag{2.3.15}$$

which is just the left-hand side of (2.3.9), and the proof is complete. □

In combination with the equivalence of the two notions of sufficiency (Theorem 7) and the Pitman-Koopman result as given by Theorem 10, an immediate consequence is the following.

Corollary 16 (Parametric conjugate models are exponential families)**.** *Let $\mathcal{M} = \{\mu_{X|\theta} | \theta \in \Omega_\theta\}$ be a parametric family and $\mathcal{N} = \{\mu_{\Theta|y} | y \in \Omega_y\}$ a family of conjugate priors. Assume that:*

1. *Ω_y is a subset of a finite-dimensional space.*
2. *$\mathcal{M}$ is dominated.*
3. *The mapping S_n is measurable for each $n \in \mathbb{N}$.*
4. *The conditional densities $f_{X|\Theta}$ of $\mathcal{M}$ have support independent of θ.*
5. *The density of the prior $\mu_{\Theta|y}$ is strictly positive on Ω_θ for all y.*

Then $\mathcal{M}$ is an exponential family model.

Condition (1) excludes, in particular, the trivial case where $\mathcal{N}$ is the set of all measure on Ω_θ. In this case,any possible posterior is necessarily an element of $\mathcal{N}$, and the model is trivially conjugate. Condition (4) may be modified according to the version of the Pitman-Koopman lemma used in the proof (cf Sec. 2.2.4 and Sec. 2.2.5). Whether the assumption of a dominated family is crucial is an interesting question, since available proofs of the Pitman-Koopman result rely on the classical definition of sufficiency, and Bayesian sufficiency is weaker in the undominated case.

Conjugacy by sufficiency

If a model admits a sufficient statistic of dimension fixed with respect to sample size, a conjugate prior of generic form can be derived from the Neyman factorization.

Theorem 17. *Let $\mathcal{M} = \{\mu_{X|\theta} | \theta \in \Omega_\theta\}$ be a dominated parametric family. Assume there is a sufficient statistic $s_n : \Omega_x^n \to \Omega_y$ such that Ω_y is contained in a finite-dimensional space. Let $g_{2,n}(s_n(x_1, \dots, x_n), \theta)$ be the corresponding s_n-dependent term in the Neyman factorization. Let ν be any measure on Ω_θ such that the denominator of*

$$g(\theta|y, n) := \frac{g_{2,n}(y, \theta)}{\int g_{2,n}(y, \theta) d\nu(\theta)} \tag{2.3.16}$$

is non-zero and finite for all $y \in \Omega_y$. Then the parametric family of densities

$$\mathcal{N} := \{g(\theta|y, n) | y \in \Omega_y, n \in \mathcal{N}\} \tag{2.3.17}$$

is conjugate to $\mathcal{M}$.

A proof is given by Schervish (1995), based on a derivation by DeGroot (1970). Pitman-Koopman theory essentially limit the applicability of the theorem to exponential family models. For an exponential family model, the result carries over to non-integer values of n. The following definition is originally due to Raiffa and Schlaifer (1961).

Definition 18 (Natural conjugate prior)**.** Let $\mathcal{M}$ be an exponential family model with sufficient statistic s. Define Ω_y as the convex hull of the image $s(\Omega_x)$. Let ν be any measure on Ω_θ such that

$$K(\lambda, y) := \int \exp(\langle \theta | y \rangle - \lambda \phi(\theta)) d\nu(\theta) \tag{2.3.18}$$

is in $(0, +\infty)$ for all $y \in \Omega_y$ and all $\lambda \in \mathbb{R}_+$. Then the family $\mathcal{N}$ defined by the conditional densities

$$g_{\Theta|\lambda,y} = \frac{1}{K(\lambda, y)} \exp(\langle \theta | y \rangle - \lambda \phi(\theta)) \ , \tag{2.3.19}$$

with respect to the measure ν is called the *natural conjugate family* of priors for $\mathcal{M}$ with respect to ν.

A family so defined is conjugate to $\mathcal{M}$ for any integer value of λ by Th. 17. Since $\mathcal{M}$ is an exponential family model, each sufficient statistic s_n is of the form $s_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n s(x_i)$, and non-integer values of λ can be derived by simple linear interpolation. By defining Ω_y as a convex hull, the image of s is closed under averaging. For any exponential family $\mathcal{M}$ with a natural conjugate family (2.3.19), the *posterior index*, i. e. the abstract mapping that specifies the posterior based on prior and observations, can be specified in closed form. If sample values $x_1, \ldots, x_n$ are observed, and the prior is by (λ, y), then the posterior is indexed by $(\tilde{\lambda}, \tilde{y})$ with

$$\tilde{\lambda} := \lambda + n \qquad \text{and} \qquad \tilde{y} := y + \sum_{i=1}^{n} s(x_i) \; . \tag{2.3.20}$$

From a technical point of view, the special form obtained for exponential family models is due to the sample-wise application of the sufficient statistic, and the log-linearity of the model in s. The consequence is a linear geometry in parameter space, where the mapped observations and their averages, parameters, and hyperparameters all constitute points in the space, and posteriors are obtained by linear interpolation. A result of Diaconis and Ylvisaker (1979) uses this linear arithmetic to characterize the set of all conjugate priors in exponential families. They show that conjugate priors are those for which the expectation of the sample mean with respect to the posterior is linear.

Theorem 19 (Diaconis-Ylvisaker characterization of conjugate priors)**.** *Let $\mathcal{M}$ be a natural exponential family dominated by Lebesgue measure. Let $\mathcal{N}$ be a family of priors. Then $\mathcal{N}$ is natural conjugate in the sense of Def. 18 if and only if*

$$\mathbb{E}_{\mu_{\Theta|X_1,\ldots,X_n}}\left[\mathbb{E}_{f_{X|\theta}}\left[X|\theta\right]\right] = \frac{y + n\hat{x}}{a+n} \; . \tag{2.3.21}$$

That is, given observations $x_1, \ldots, x_n$, the expected value of a new draw x under unknown value of the parameter is linear in the sample average $\hat{x} = \frac{1}{n}\sum x_i$.

Interpretation of Natural Conjugate Priors

A natural conjugate prior of a given exponential family model is the model equivalent to that family under the transformation defined by the sufficient statistic. Consequently, the prior can be interpreted as a posterior of the same model for “initial” or “previous” observations drawn under a uniform prior. The prior parameters λ and y can then be interpreted as the initial sample size and the average sufficient statistic of the initial sample, respectively.

This is easily made precise for the case where s has a differentiable inverse. As a sufficient statistic is applied pointwise, it can be regarded as a preprocessing step. Define a new random variable $U := s(X)$, and assume that any observations $x_1, x_2, \ldots$ are put through the map s, yielding $u_1, u_2, \ldots$, and then are forgotten. Accordingly, the model (2.2.15) may be translated into an equivalent model on Ω_θ, by transforming out the sufficient statistic s: If s is invertible, then $x(u) = s^{-1}(u)$ is well-defined, and

$$f(x)dx = f(x(u))\frac{dx}{du}du = f(s^{-1}(u))\frac{ds^{-1}}{du}(u)du \; . \tag{2.3.22}$$

In particular, for the conditional density (2.2.15),

$$f_{X|\theta}(x|\theta)dx = f_{X|\theta}(s^{-1}(u)|\theta)\frac{ds^{-1}}{du}(u)du = \tilde{h}(u)\exp\left(\langle u|\theta\rangle - \phi(\theta)\right)du \; , \tag{2.3.23}$$

where $\tilde{h}(u) := h(s^{-1}(u))\frac{ds^{-1}}{du}(u)$. For n observations, the density of the average $\hat{u}_n$ is, by n-fold convolution of the density in (2.3.23) with itself, given by $\left(*_{i=1}^n h\right)(\hat{u}_n)\exp\left(\langle \hat{u}_n|\theta\rangle - n\phi(\theta)\right)$. Absorbing the convolution of h into the carrier, $\hat{u}_n$ has density

$$p(\theta|\hat{u}_n) \propto \exp\left(\langle \hat{u}_n|\theta\rangle - n\phi(\theta)\right) \; , \tag{2.3.24}$$

which is precisely the canonical conjugate prior family of $\mathcal{M}$, with $y = \hat{u}_n$ and $\lambda = n$. In short, the canonical conjugate prior can be regarded as a result of the following program:

1. Choose a uniform prior $U(\theta)$ on the parameter domain Ω_θ. Depending on the domain, U may or may not be proper.
2. Draw n initial samples $x_1, \ldots, x_n$ from the Bayesian model.
3. Compute the average $\hat{u}_n$ of the sufficient statistic.
4. Define the prior on θ as the posterior of θ given $\hat{u}_n$ under the uniform prior.

2.4 Bayesian Nonparametrics

The models commonly referred to as *nonparametric Bayesian* were originally introduced to apply Bayesian techniques in a manner similar to classical nonparametric methods, which allow the number of parameters or explanatory variables to grow with sample size (Ferguson, 1973). The result has been a class of Bayesian models on infinite-dimensional spaces, which includes the Dirichlet process and Polya trees as prominent examples, and terminology has been extended in hindsight to include other infinite-dimensional models, in particular Gaussian process priors. A distinguished role among nonparametric Bayesian models play those which define distributions on distributions, and are applicable as priors in Bayesian problems. Such priors do not randomly generate a parameter for the sampling distribution, but instead generate the sampling distribution itself, though there is no sharp distinction between the two cases for infinite-dimensional parameter sets. Most statistics texts on nonparametric Bayesian models exclusively discuss random distribution models. Overviews along these lines include Ferguson (1974); Walker *et al.* (1999); Ghosh and Ramamoorthi (2002); Müller and Quintana (2004). For an introduction to Gaussian process models, see for example Rasmussen and Williams (2006).

2.4.1 Basic Definition

Parametric estimation techniques are methods which (i) estimate a parametric model from data and for which (ii) the dimension of parameter space is constantly upper-bounded w. r. t. sample size. Roughly speaking, for asymptotic sample size, the definition assures an infinite number of observations to be available per parameter dimension. *Nonparametric* estimation methods are methods which do not require (ii). The classic example of nonparametric methods are Parzen or kernel density estimators. These models are parameterized, by a global bandwidth and one location parameter per observation, but are "nonparametric" because the number of parameters grows with sample size.

Bayesian models do not generally tie in well with nonparametric strategies. They are inherently parametric, because they define a prior probability on a given parameter space. Changing the parameter space by adding dimensions amounts to switching models. For example, Parzen estimators center a Gaussian at each observation. They may, in principle, be equipped with a prior on bandwidth or location, but each sample would require its own posterior, and the Bayesian model would have to replicate n-fold. Moreover, the model would require a conceptual reinterpretation, since each individual Gaussian density is used as a smoothing kernel, rather than a probability model of the generative process explaining the sample. So-called *Bayesian nonparametric* models approach the problem by providing a large number of parameters, only a few of which are used per sample observation. Several introductory texts (e. g. Schervish, 1995) characterize such models as Bayesian models on infinite-dimensional spaces. Here is a slightly different definition:

Definition 20 (Nonparametric Bayesian model)**.** A Bayesian model, consisting of a parametric sampling model $\mu_X(X|\Theta)$ and a prior distribution $\mu_\Theta(\Theta)$, is called nonparametric if (i) there is a number $n_0 \in \mathbb{N}$ such that explaining each additional observation requires at most n_0 additional degrees of freedom in parameter space, and (ii) the expected number of degrees of freedom required to explain any observation $x_1, \dots, x_n$ is monotonically increasing in n.

An obviously necessary provision not made here explicitly is that the model has a sufficient number of degrees of freedom available to explain a given sample, which is the principal motivation for defining models of infinite dimension. If a nonparametric Bayesian model has an infinite number of degrees of freedom, it can explain a sample of any given finite size. Taking an infinite limit of the sample size is also possible without changing the model, so asymptotic behavior of the model can be studied. Therefore, almost all nonparametric Bayesian models common in the literature are infinite-dimensional models, and the terms "nonparametric Bayesian model" and "infinite-dimensional Bayesian model" are used equivalently by some authors. All nonparametric Bayesian models considered in the following chapters are also of infinite-dimensional type.

Nonetheless, it is worth noting that the characteristic property of a nonparametric Bayesian model is not infinite dimensionality, but the ability to explain partial observations. A model of high, but finite dimension may be perfectly sufficient if sample size is bounded in advance. Required are a rule for how to explain individual observations by means of some of the d degrees of freedom, and how to choose d given the sample size. The following definition formalizes the idea of a partial observation, a sample that accounts only for a subset of the model's degrees of freedom.

Definition 21 (Partial observation)**.** Let X be a random variable with multiple degrees of freedom, i. e. with values in a space Ω^{E} of product structure

$$\Omega^{\mathrm{E}} = \prod_{i \in E} \Omega^{\{i\}} , \tag{2.4.1}$$

where $\Omega^{\{i\}}$ are arbitrary component spaces. For any $I \in E$, the respective partial product over elements of I only will be denoted Ω^{I}. Then an observation of the restricted variable $X^{\mathrm{I}} = X|_{\Omega^{\mathrm{I}}}$ will be called a *partial observation*[2].

[2] The concept of a partial observation may be regarded as a form of censored data. The term partial is used here instead, because "censoring" is typically taken to imply a *systematic* effect, such as right-censoring or interval-censoring. For a partial observation X^{I}, the selection of indices I at which measurements are available may itself be random.

In a nonparametric Bayesian model, the partial observation X^{I} usually represents the finite observed sample $x_1, \ldots, x_n$, and the size of the index set I grows with n. Bayesian and classical nonparametric models differ in how the estimation process affects model dimension, Classical models discard superfluous dimensions not determined by data. Bayesian nonparametric models keep all dimensions, and determine degrees of freedom that cannot be estimated from data by prior assumption. A Parzen estimate for a sample of size n has precisely n location parameters. A nonparametric Bayesian model of parameter dimension d, where d may or may not be finite, will estimate a d-dimensional posterior regardless of sample size.

2.4.2 Construction Techniques

If nonparametric Bayesian methods require measures on infinite-dimensional spaces, the first question to consider is whether and how such measures can be defined, and represented in a manner suitable for inference. A Gaussian of finite dimension can be written in closed form, as a density with respect to Lebesgue measure. Extending the concept of a Gaussian to infinite-dimensional space is less straightforward. There is no such thing as a meaningful "infinite-dimensional limit" of the density function, because its carrier, Lebesgue measure, cannot be extended to infinite-dimensional space (Skorohod, 1974). Certain interesting infinite-dimensional models which have been constructed in other ways can be shown not to admit density representations (including the Dirichlet process). A necessary step in nonparametric Bayesian constructions is thus to abandon the familiar notion of modeling with densities, and look for alternative representations. Since the introduction of the Dirichlet process in 1973, various constructions have been considered, from among which the following have emerged:

1. Modified stochastic processes (constructive)
2. Subdivision strategies (constructive)
3. De Finetti's theorem (unconstructive)
4. Kolmogorov's extension theorem (constructive)

All four approaches will be discussed in more detail in the following, with a particular emphasis on (4). They are not mutually exclusive, and the Dirichlet process in particular can be derived by means of each. The extension theorem is arguably the most powerful approach, since any of the processes in (1) and (2) can also be constructed by (4), even if the approach of modifying an existing process with known properties or of applying a subdivision method may be more convenient.

Modified stochastic processes

This approach is generally popular for the generation of random probability distributions over the real line or its intervals. The distribution is defined by drawing the cumulative distribution function as the sample path of a stochastic process with non-negative increments. For example, Ferguson (1973) gives a definition of the DP on an interval $[a, b]$ which generates a CDF as follows: (1) generate a random function f as the sample path of a gamma process on $[a, b]$, and (2) normalize f by setting $\bar{f}(x) := \frac{f(x)}{f(b)}$. The use of CDFs largely restricts the approach to the real line. In principle, one may consider generating a density function instead, but the merit of CDFs is that their properties apart from normalization, i. e. a zero limit at the lower interval boundary and monotonicity, are local. Local properties can be guaranteed a. s. by conditions on the increments of a process (the positive increments of the gamma process guarantee a. s. monotonicity of its trajectory).

A rather general class of such processes that has emerged as priors on CDFs are *neutral to the right* (NTR) processes Doksum (1974); Ferguson and Phadia (1979). Lévy processes are another class of candidate measures, as the most thoroughly studied type of independent increment processes. Their application in Bayesian nonparametrics has been considered by Wolpert *et al.* (2003); Wolpert and Ickstadt (2004). Lévy processes have a certain practical appeal, because of the simple closed form of their characteristic function. The point here is that parameterized classes of stochastic processes do not in general admit a joint representation in form of a conditional density, as standard parametric models do. But even measures without a density still have a well-defined characteristic function (Fourier transform), and for some processes, the characteristic function has a simple form depending on the process parameters. This is the case, for example, for Gaussian processes, for which the characteristic function has a functional form closely resembling the Gaussian density in finite dimensions (Skorohod, 1974). The same is true for Lévy processes, by the Lévy-Khinchine formula. For such models, the characteristic function does not just provide an abstract tool for convergence proofs, but substitute (to some degree) for a density representation. At least for practical applications, where a tractable posterior is required, the idea of generating a CDF as a normalized draw from a Lévy process turns out to be of limited scope. To generate a CDF, a Lévy process must be non-decreasing (a so-called *subordinator*). The only such model that admits a conjugate posterior after normalization is the gamma process (James *et al.*, 2005). The normalized model is then, once again, the DP. Other common types of stochastic processes remain to be studied in detail. Processes defined by means of stochastic differential equations have been studied for Bayesian nonparametric modeling e. g. in recent work of Griffin (2007). The "exponential families of stochastic processes" of Küchler and Sørensen (1997) may be another promising class of models. These are parameterized stochastic processes with a parameter function θ and a time index t. The definition of the exponential family stochastic process model requires that for all $t < \infty$, the process has a likelihood function of the form

$$L^t(\theta) = \frac{1}{Z^t(\theta)} \exp\Big(\sum_{i=1}^{d} s_i^t(\theta) X_i^t\Big) , \tag{2.4.2}$$

where (i) d is independent of the time index t and (ii) the components $X_1^t, \ldots, X_d^t$ are real-valued stochastic processes, which may depend on the past, but not on the future. These models may be of particular interest for Bayesian nonparametrics because they unify counting processes, autoregressive models, diffusions and jump-diffusions, random fields and Lévy processes from an exponential family point of view.

Subdivision Strategies

A random probability measure is a random function on a σ-algebra (with some special properties). σ-algebras can be generated from systems of partitions, and can be constructed as a consistent rule for assigning probability mass to random partitions of its domain into measurable sets. Subdivision or random partition constructions emphasize this point of view. The "Chinese restaurant process" was constructed in this manner by L. E. Dubins and J. Pitman (Pitman, 1995). Based on work of Doksum (1974) and Ferguson (1974) on tailfree processes, Mauldin *et al.* (1992) suggest to generate random probabilities over partitions of the domain by means of a tree. The definition of the random measure requires a *given* recursive partitioning of the domain, i. e. a set of partitions $\mathcal{H}^n$ such that $\mathcal{H}^{n+1}$ is a refinement of $\mathcal{H}^n$. (The σ-algebras generated by each $\mathcal{H}^n$ form a filtration indexed by n.) It is no loss of generality to assume that for $n \mapsto n+1$, each set in $\mathcal{H}^n$ is partitioned into two subsets. The partitions can then be arranged in a binary tree, and probabilities are assigned in a consistent manner by equipping each node with a probability for its right and left subtree. The probability assigned to a set on level n is the product over all probabilities along the path from the root to the set node. The measure so obtained becomes random if the edge probabilities are drawn at random. Not surprisingly, the construction assumes these probabilities to be Dirichlet random vectors (with two entries, and therefore beta random variables). The model is called a *Pólya tree*, and has been studied with some attention because it contains the DP as a special case, and for other settings of the parameters can generate almost surely continuous random measures.

De Finetti's Theorem

An exchangeable sequence of random variables (with values in Polish spaces) is conditionally independent, by de Finetti's theorem (Th. 48 in App. A). The joint measure of the sequence can be represented as the mixture of a product measure against a suitable mixing distribution (sometimes called the "de Finetti prior"). Hewitt and Savage (1955) have shown that for a given sequence, the mixing distribution is unique. Defining an infinite sequence of exchangeable random variables, for example by specifying a generation algorithm that guarantees exchangeability, thus implicitly but uniquely defines a measure. In principle, this is a construction approach, though not a constructive one. Blackwell and MacQueen (1973) construct the DP by proving that it is the mixing distribution for the generative model which they call the "infinite Polya urn scheme". De Finetti's theorem is given, for reference, in Sec. A. Its application in the Bayesian context is discussed by Bernardo and Smith (1994). Kallenberg (2005) provides a thorough treatise in terms of probability theory.

Kolmogorov's Extension Theorem

The extension theorem constructs a measure on an infinite-dimensional space directly from its finite-dimensional marginals. The customary textbook definition of the Gaussian process, for example, defines it as "a collection of random variables, any finite number of which have a joint Gaussian distribution". The Kolmogorov theorem guarantees that, given these joint Gaussian distributions on all possible finite subsets of the collection, the Gaussian process measure on the whole (usually infinite) set exists and is unique. Similarly, Ferguson (1973) defines the DP on general domains by means of the extension theorem, as the infinite-dimensional measure defined by Dirichlet marginals. The discussion in Ch. 5 is based primarily on the extension theorem, and so the next paragraph will describe the technique in some detail.

2.4.3 Kolmogorov's Extension Theorem

The extension theorem of Kolmogorov provides the most general tool available for the construction of nonparametric Bayesian models. It is not usually required for the discussion of standard models such as the Dirichlet, but is presented here in some detail because of its substantial role in Ch. 5.

Construction of Infinite-Dimensional Probability Models

Bayesian nonparametric models are typically probability distributions on some random object with an infinite number of elements, and the first problem to consider in Bayesian nonparametrics is how such a distribution can be defined. More precisely, denote the individual, "scalar" elements of the random variable by $X^{\{i\}}$, where $i \in E$ and E is an infinite index set. The elements can be collected to form an infinite quantity X^{E}, which may be an infinite vector, a function, an operator, an infinite graph etc. The question addressed in this paragraph is: If the distributions of the individual elements $X^{\{i\}}$ are known, and possibly the joint distributions on finite subsets of elements, how can a joint distribution on X^{E} be specified?

Construction methods for probability measures on infinite-dimensional objects have been thoroughly studied in the theory of stochastic processes. Two techniques are of fundamental importance:

1. Product measure constructions: Independent variables or increments.
2. Kolmogorov's extension theorem: Measures with dependency structures.

The first case is rather obvious: Say measures $\mu^{\{i\}}$ are known. Then these can be combined to form an infinite product measure

$$\mu := \bigotimes_{i \in E} \mu^{\{i\}} . \tag{2.4.3}$$

If each random variable has sample space $\Omega^{\{i\}}$, and is defined on a σ-algebra $\mathcal{A}^{\{i\}}$, then the product measure lives on space $\prod_{i \in E} \Omega^{\{i\}}$ with σ-algebra $\bigotimes_{i \in E} \mathcal{A}^{\{i\}}$. No further assumptions are required for μ to be well-defined, which is worth noting, because some topological structure *will* be required if the random variables are to be dependent. Apparently, though, the construction is not particularly interesting; after all, why construct a joint measure if it treats random variables individually anyway. However, measures on independent random quantities are put to great use in stochastic process theory by interpreting the elements as increments rather than variables of a process. The resulting *independent increment* processes constitute a major part of the stochastic process landscape, and also find applications in Bayesian nonparametrics.

The second means of construction is Kolmogorov's extension theorem. In a nutshell, the theorem states that the joint distribution on X is completely determined if all its marginals on *finite* subsets $\{X^{\{i\}} | i \in I, I \text{ finite}\}$ of axes are known. Somewhat surprisingly, the marginals on finite-dimensional subsets suffice. The Kolmogorov theorem is a regularity result, stating that a set function which satisfies the requirements of a probability measure cannot be arbitrarily complex, and hence knowing those of its values which are determined by the marginals are sufficient to completely determine the measure.

Mathematical setting

Construction results for product measures (independent random variables) can be proven under the sole assumption that the given components are probability measures on arbitrary σ-algebras. Considering random variables of arbitrary dependence structure requires certain minimal conditions on the topology of the underlying spaces. The general case that has been established are random variables that take values in so-called *Polish product spaces*, equipped with a Borel σ-algebra. A Polish space is a separable, metrizable topological space.[3] Borel σ-algebras are those generated by the open sets, and hence defined by the topology of a space. Intuitively, the definition of a Polish space is chosen to ensure that both the space (the underlying point set) and its σ-algebra behave sufficiently similar to the real numbers for essential properties of measures to be preserved. Examples of Polish spaces include Euclidean space, any separable Banach or Hilbert space, countable discrete spaces, countable products of spaces that are themselves Polish, and any open or closed subset of another Polish space.

Spaces and σ-algebras. We again consider the infinite set of random variables $X^{\{i\}}$, indexed by the infinite set E. For illustration, suppose the construction defines an Gaussian process, i. e. an infinite-dimensional Gaussian distribution on the variable X^{E}. A random value drawn from this distribution can be thought of as an infinite-dimensional vector, and hence as a random function. Then the index set E of dimensions is the domain of the function. We thus have to choose $E = \mathbb{R}$, or $E = [0,1]^3$ or $E = \mathbb{Z}$

[3] Kolmogorov originally formulated the extension theorem for products of real axes (Euclidean spaces and their infinite limit). Products of Polish spaces are a considerable generalization. They are not required to be vector spaces, and hence have no a priorily attached notion of an algebraic operation on their elements, nor need they satisfy the same closure properties. Being metrizable means the topology is induced by some metric. The definition of the space does not imply any particular metric, but the condition ensures that the structure of its topology is consistent with the properties of metrics. Separability is the existence of a countable number of open sets, such that every open set is representable as a union of some of these sets. The open sets generate the σ-algebra, and σ-algebras are closed under countable unions. The separability condition ensures that, based on a countable system of sets, a σ-algebra can be defined which makes every open set measurable. For metrizable spaces, an equivalent criterion for separability is that any point can be approximated arbitrarily well by the elements of a countable set, which is just the case if the space has a dense countable subset. For reference see Fremlin (2006); Bourbaki (2003).

to obtain random functions on the real line, or on the three-dimensional unit cube, or on the integers, respectively. Formally, the only requirement on E is that it is non-empty. Denote by $E^* = \{I \subset E || I| < \infty\}$ the set of its finite subsets. All individual component variables $X^{\{i\}}$ take values in a Polish space $\Omega^{\{i\}} = \Omega$. The sample space of the infinite-dimensional random variable is the infinite product space $\Omega^{\mathrm{E}} = \prod_{i \in E} \Omega$. A notable difference to the construction of product measures mentioned above is that Ω^{E} is an infinite repetition of the same space Ω. The finite-dimensional marginals, used to define the infinite-dimensional measure, are the marginal distributions on the finite-dimensional subspaces of Ω^{E}. Any finite subset $I \in E^*$ of indices determines such a subspace, which will be denoted Ω^{I}. Each marginal random variable X^{I} has a marginal measure μ^{I}, defined on the Borel algebra $\mathcal{B}^{\mathrm{I}}$ on Ω^{I}. The domain of the constructed, infinite-dimensional measure will be the infinite product algebra $\mathcal{B}^{\mathrm{E}} = \bigotimes_{i \in E} \mathcal{B}(\Omega)$.

Projections and marginals. A key notion in the following will be that of a *projection*, because marginals are projections. The equivalence of marginals and projections is due to the product structure of the space, and is of great convenience, because a probabilistic operation (marginalization) can be represented as a geometric one (projection). A projection operator between pairs of subspaces of Ω^{E} is defined as follows: Let $I \subset J$. Then the projector $\mathrm{P}_{\mathrm{J,I}}$ as the mapping which takes each element of Ω^{J} to its restriction on I. That is, if $x^{\mathrm{J}} \in \Omega^{\mathrm{J}}$ is regarded as a list $(x_i)_{i \in J}$, the projector removes all elements except those with index in I, $\mathrm{P}_{\mathrm{J,I}} x^{\mathrm{J}} = (x_i)_{i \in I}$. The projection immediately generalizes to sets (by application to all points in the set). The *preimage under projection* by $\mathrm{P}_{\mathrm{J,I}}$ will be denoted $\mathrm{R}_{\mathrm{J,I}}$, that is, $\mathrm{R}_{\mathrm{J,I}} x^{\mathrm{I}} = \{x^{\mathrm{J}} \in \Omega^{\mathrm{J}} | \mathrm{P}_{\mathrm{J,I}} x^{\mathrm{J}} = x^{\mathrm{I}}\}$. For obvious reasons, preimages under projection are often referred to in the literature as *cylinder sets* (imagine the preimage under projection of a disc-shaped set). If $A^{\mathrm{I}} \subset \Omega^{\mathrm{I}}$, the preimage $\mathrm{R}_{\mathrm{E,I}} A^{\mathrm{I}}$ is called the *cylinder with base* A^{I}. For a given space Ω^{I} with $I \in E^*$, the set of all cylinders with base in the σ-algebra $\mathcal{B}^{\mathrm{I}}$, the system $\mathrm{R}_{\mathrm{E,I}} \mathcal{B}^{\mathrm{I}}$, is again a σ-algebra. The union $\mathcal{Z}(\Omega) = \bigcup_{I \in E^*} \mathrm{R}_{\mathrm{E,I}} \mathcal{B}^{\mathrm{I}}$ forms an algebra[4], but no σ-algebra.

The Extension Theorem

The pivotal ingredient of the extension theorem is the following definition:

[4] Like a σ-algebra, an algebra is a system of a set, with almost the same properties as a σ-algebra, except for closure under intersection: An algebra is only required to be closed under intersections of any finite number of sets (whereas a σ-algebra must be closed under countably infinite intersections).

Definition 22 (Projective family)**.** Let $\{\mu^{\mathrm{I}}|I \in E^*\}$ be a family of probability measures on the spaces $(\Omega^{\mathrm{I}}, \mathcal{B}^{\mathrm{I}})$. The family is called a *projective family* if, for any $I, J \in E^*$ with $I \subset J$,

$$\mathrm{P}_{\mathrm{J,I}}\mu^{\mathrm{J}} = \mu^{\mathrm{I}} \ , \tag{2.4.4}$$

or equivalently, $\mu^{\mathrm{J}}(\mathrm{R}_{\mathrm{J,I}}A^{\mathrm{I}}) = \mu^{\mathrm{I}}(A^{\mathrm{I}})$ for any $A^{\mathrm{I}} \in \mathcal{B}^{\mathrm{I}}$.

Suppose the measure μ^{E} of the infinite variable X^{E} was already given. Then, if all its marginals were computed on the finite-dimensional subspaces of Ω^{E}, these marginals would be consistent in the following sense: Say $J, K \in E^*$ are two sets of axes which overlap. Let I be a common subset, $I \subset J$ and $I \subset K$. Then the marginals of μ^{J} and μ^{K} on the common subspace Ω^{I} must be identical. If marginalization is formalized as projection, the resulting relation between marginals is just Eq. 2.4.4. In other words, the definition above states that a projective family is a system of measures that *could* form the marginals of a common measure μ^{E}, if such a measure exists. The Kolmogorov theorem states that the downward projection is reversible: If the measures are projective, the measure μ^{E} exists and is unique.

Theorem 23 (Kolmogorov extension theorem)**.** *Let $\{\mu^{I}|I \in E^*\}$ be a projective family of probability measures on the spaces $(\Omega^{I}, \mathcal{B}^{I})$. Then there exists a uniquely defined measure μ^{E} on $(\Omega^{E}, \mathcal{B}^{E})$ with the measures μ^{I} as its marginals.*

The measure μ defined by extension is called the *projective limit* of the projective family $\{\mu^{\mathrm{I}}|I \in E^*\}$. The intuitive meaning of the theorem is roughly the following: A probability measure is a set function of the form $\mu : \mathcal{A} \to [0, 1]$, where $\mathcal{A}$ is a σ-algebra. To satisfy the definition of a probability measure, an arbitrary set function has to satisfy a number of conditions. The conditions impose an amount of regularity on μ that severely restricts its degrees of freedom. As a consequence, if a set function is a probability measure, it can be completely determined by its values on a suitable subset of the domain $\mathcal{B}^{\mathrm{E}}$. This is roughly comparable to other forms of regularity, such as a continuous function being completely determined by its values on a dense subset, or band-limited functions being determined by their values on a grid. The extension theorem states that the subset $\mathcal{Z}$ of $\mathcal{B}^{\mathrm{E}}$ (on which the values of μ^{E} are given by the marginals) is sufficiently rich to completely determine μ^{E} on the whole of $\mathcal{B}^{\mathrm{E}}$.

Remark 24 (Extension to uncountable dimensions)**.** The extension theorem holds irrespectively of whether or not the index set E is countable, but the uncountable case may lead to certain complications. The theorem defines a measure on the infinite product algebra $\mathcal{B}(\Omega)^{\mathrm{E}}$. The domain of interest is usually the Borel algebra of the infinite-dimensional product space $\mathcal{B}(\Omega^{\mathrm{E}})$. If E is countable, the two are identical, $\mathcal{B}(\Omega)^{\mathrm{E}} = \mathcal{B}(\Omega^{\mathrm{E}})$. If E is not countable, then generally only $\mathcal{B}(\Omega)^{\mathrm{E}} \subset \mathcal{B}(\Omega^{\mathrm{E}})$, such that the extension theorem defines a measure only on part of the domain of interest. In particular, if Ω contains more than one element, the singletons of Ω^{E} (the subsets of Ω^{E} containing only a single point) are not in $\mathcal{B}(\Omega)^{\mathrm{E}}$. In other words, the extended measure, which assigns values to sets included in $\mathcal{B}(\Omega)^{\mathrm{E}}$, cannot be applied to sets of the form $\{x^{\mathrm{E}}\}$, where $x^{\mathrm{E}} \in \Omega^{\mathrm{E}}$. Intuitively, the extended measure and σ-algebra constructed by the extension theorem are too coarse to resolve the singletons. For Bayesian estimation, sample observations are singletons, and should be measurable. Hence for some examples, notably the Dirichlet process, additional considerations are required when defining the domain of the process measure, such as restriction of the continuous domain to a dense, countable subset.

2.4.4 Dirichlet Processes

All construction approaches sketched above yield the DP as a common special case, and so a variety of definitions is available. Definition 25 below is due to Ferguson (1973), and based on Kolmogorov's extension theorem.

Definition and Basic Properties

The Dirichlet process is, roughly speaking, the infinite extension of the finite-dimensional Dirichlet distribution. More precisely, it is the projective limit of a projective family of finite-dimensional Dirichlet distributions, as defined by Th. 23. The Dirichlet distribution is an exponential family model on the d-dimensional real simplex $\mathrm{Sim}\,(\mathbb{R}, d)$, which generates finite probability distributions on d events. As an exponential family model, it has a concentration parameter $\beta \in \mathbb{R}_+$ and an expectation parameter π. Since the simplex is closed under averaging, the expectation is also a finite probability distribution, $\pi \in \mathrm{Sim}\,(\mathbb{R}, d)$. The density is

$$p_{\mathrm{Dir}}(\theta|\beta,\pi) := \frac{1}{Z_{\mathrm{Dir}}(\beta,\pi)} \exp\Big(\sum_{j=1}^{d} (\beta\pi_j - 1)\log\theta_j\Big) \tag{2.4.5}$$

with partition function $Z_{\mathrm{Dir}}(\beta,\pi) := \Gamma(\beta)^{-1}\prod_{j=1}^{d}\Gamma(\beta\pi_j)$. In terms of exponential family models, the parameter and sample space are identically $\mathrm{Sim}\,(\mathbb{R}, d)$, and the sufficient statistic is the identity mapping[5]. The model

[5] The name "Dirichlet distribution" is due to the partition function, rather than the actual probability model. The partition function is an integral over the simplex, and thus

is the natural conjugate prior of the multinomial family. To apply the extension theorem, we have to choose a suitable index set and subset structure. The Gaussian process construction uses index sets $I \in E^*$ consisting of a finite number of points in the domain E of the random function. The I-marginal of the Gaussian process represents the distribution of the random function's values at these points. The Dirichlet process construction attempts to produce random probability measures on some sample space Ω. The domain of a measure is a σ-algebra. In analogy to the GP case, the index set of axes E should therefore be a σ-algebra. This idea is slightly modified for the construction of the DP from Dirichlet distribution marginals. Random draws from a Dirichlet distribution are finite probability distributions, and their domain is a special form of σ-algebra, generated by partitioning the sample space Ω into a finite number "histogram bins". The finite index subsets $I \in E^*$, which corresponded to finite sets of points in the Gaussian case, will now represent partitions of Ω into a finite number of subsets: For a measurable space $(\Omega, \mathcal{A})$, a subdivision of Ω into measurable sets will be called a *measurable partition*. That is, $H = (A_1, \dots, A_n)$ is a measurable partition if $A_i \in \mathcal{A}$ for all i, $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^n A_i = \Omega$. The system of all (possibly infinite) $\mathcal{A}$-measurable partitions will be denoted by $\mathcal{H}$, and by $\mathcal{H}^*$ the subsystem of those partitions consisting only of a finite number of sets each.

Definition 25 (Dirichlet process)**.** Let $(\Omega, \mathcal{A})$ be a measurable space, with a probability measure G_0. For any partition $H \in \mathcal{H}^*(\mathcal{A})$, let $\mathbb{R}^{\mathrm{H}}$ be the product space

$$\mathbb{R}^{\mathrm{H}} := \prod_{A \in H} \mathbb{R} \,, \tag{2.4.6}$$

and $\mathrm{Sim}\,(\mathbb{R}, H)$ its unit simplex. Denote by $p^{\mathrm{H}}_{\mathrm{Dir}}(\,.\,|\alpha, g)$ the Dirichlet density on $\mathrm{Sim}\,(\mathbb{R}, H)$, with concentration $\alpha \in \mathbb{R}_+$ and expectation $g \in \mathrm{Sim}\,(\mathbb{R}, H)$. For each $H \in \mathcal{H}(\mathcal{B})$, define the vector $g^{\mathrm{H}} \in \mathrm{Sim}\,(\mathbb{R}, H)$ by

$$\forall A_i \in H : \qquad g^{\mathrm{H}}_i := G_0(A) \,. \tag{2.4.7}$$

Denote by μ^{H} the measure specified by the density $p^{\mathrm{H}}_{\mathrm{Dir}}(\,.\,|\alpha, g^{\mathrm{H}})$ (with respect to Lebesgue measure on the respective simplex). Then the projective limit

over a domain defined by inequality constraints ($\theta_i > 0$ for all i) and equality constraints ($\sum_i \theta_i = 1$). Such integrals were first studied for the case of integration over a sphere by Dirichlet in his work on partial differential equations, and came to be called *Dirichlet integrals*. They are, in a sense, the integration counterpart of Lagrange optimization problems, which restrict the differentiation (rather than integration) problem to a set defined by algebraic constraints.

of the projective family $\{\mu^{\mathrm{H}}|H \in \mathcal{H}(\mathcal{B})\}$ is called a *Dirichlet process*[6] with *base measure* G_0, and will be denoted DP (α, G_0).

The key to this construction is the way marginalization works for Dirichlet distributions. In Gaussian models, for example, marginalization is deletion. A dimension is marginalized out by deleting the corresponding entries from random vectors and parameters. In a Dirichlet model, marginalization is combination: The Dirichlet is a distribution over histogram bins, and each entry in a Dirichlet vector, either random or parameter vector, is the probability of one bin. To remove a dimension, two bins are combined, and their entries are added. Decreasing the dimension by marginalization means coarsening the bin resolution on the domain. If $H_1, H_2 \in \mathcal{H}^*(\mathcal{A})$ are partitions such that H_2 is a refinement of H_1, then $p^{\mathrm{H}_2}_{\mathrm{Dir}}(\,.\,|\alpha, g^{\mathrm{H}_2})$ is an $|H_2|$-dimensional Dirichlet distribution, and $p^{\mathrm{H}_1}_{\mathrm{Dir}}(\,.\,|\alpha, g^{\mathrm{H}_1})$ is its marginal on an $|H_1|$-dimensional subspace.

Properties of the Dirichlet Process

Most essential properties of the DP are direct consequences of the properties of the Dirichlet distribution, and the presentation below emphasizes this point of view. The presentation is purely heuristic, but will be made more precise in Ch. 5. Most DP properties mentioned below are proven by Ferguson (1973), though he does not argue in terms of the marginals.

Conjugacy. The Dirichlet distribution is the natural conjugate prior of the multinomial distribution, with density

$$p_{\mathrm{Mult}}(h|\theta) = \frac{(\sum_j h_j)!}{\prod_j h_j!} \exp\Big(\sum_{j=1}^{d} h_j \log \theta_j\Big)\ . \tag{2.4.8}$$

The multinomial generates histograms h with a given number of observations. The conjugacy of the two distributions extends to the infinite-dimensional case, if the projective limit of the multinomial is taken along with that of the Dirichlet.

[6] Strictly speaking, the measure so defined lives on the product space $\mathbb{R}^{\mathcal{H}}$ and its cylinder algebra. A rigorous definition of the Dirichlet process requires, as a second step, restriction of the projective limit measure to an equivalent stochastic process (a process with identical marginals) on the subspace formed by the probability measures on Ω, which is *not* measurable in the cylinder algebra. Such a restriction (modification) can be shown to exist, according to a well-known theorem of Doob (1953), by verifying that the set of measures has outer measure one under the projective limit process on $\mathbb{R}^{\mathcal{H}}$. Ferguson (1973) neglects this point, without consequence for his further results (which are all proven under the assumption that the projective limit lives on the set of probability measures). Ghosh and Ramamoorthi (2002) point out that the projective is actually a measure on $\mathbb{R}^{\mathcal{H}}$.

The projective limit of the multinomial model is an infinite-dimensional multinomial process, which is equivalent to the "Chinese restaurant process" of Dubins and Pitman (see Pitman, 1995). These two processes are conjugate if their marginals are (cf Ch. 5). In the finite-dimensional case, if h is an observed histogram, then by the properties of exponential families, the posterior is $p_{\text{Dir}}(\theta|\beta+1, \pi+h)$. The vector π can be regarded as a function of its index, mapping $j \mapsto \pi_j$. If h contains only a single observation, in bin j_0, then $\pi + h$ is equivalent to the function $j \mapsto \pi_j + \delta_{j,j_0}$. In the limit of infinitely small bins, the indices j are replaced by elements $x \in \Omega$, and π by the measure G_0. A single observation x_0 in the domain Ω is an element of the singleton bin $\{x_0\}$. The function of j is then substituted by $x \mapsto G_0(x) + \delta_{x_0}(x)$. The posterior under observation of an "infinite histogram" with a single count, i. e. of a value in Ω, is thus a Dirichlet process with expectation $\alpha G_0 + \delta_{x_0}$ (Ferguson, 1973).

Sampling. Consider a d-dimensional Dirichlet-multinomial model, with Dirichlet prior $p_{\text{Dir}}(\theta|\beta, \pi)$. Assume that a histogram h containing a single observation is drawn from the multinomial in a two-stage manner, by drawing θ from the prior and h from the multinomial parameterized by θ. If θ is not observed, it can be integrated out, and h is drawn from the expectation π of the prior. The resulting behavior for the projective limit DP is that, if a random probability measure $G \sim \text{DP}(\alpha G_0)$, and $x \sim G$, then by integrating out G, $x \sim G_0$. Multiple observations $x_1, \dots, x_n$, assumed to be drawn from the same random G, can be generated by iterating the argument: Draw each x_{i+1} from the posterior under $x_1, \dots, x_i$. The result is the well-known DP sampling formula

$$x_{i+1}|x_1, \dots, x_i \sim \alpha G_0 + \sum_{j=1}^{i} \delta_{x_j} \ . \tag{2.4.9}$$

The formula interpolates prior assumption and data in the manner common to all exponential family models (cf. Sec. 2.3.2).

Concentration behavior. By the properties of exponential family models, the finite probability distribution π is the expected value of the distribution, and β determines the concentration (large β means tight concentration around the expected value). If π is uniform and $\beta = 1$, the Dirichlet becomes uniform on the simplex. If β is chosen close to zero, the distribution concentrates its mass at points far away from the expected value. For π uniform (i. e. at the center of the simplex), the far-away points are just the extremal points (the corners) of the simplex. The behavior carries over to the limit: In exponential family models, a large concentration parameter lets the measure concentrate tightly around its expected value. If α is large, sampling according to Eq. (2.4.9) will for most draws result in a draw from G_0, such that the overall empirical distribution converges to G_0. Since the data term (the final sum in Eq. (2.4.9)) gains in relative weight as the number of samples increases, the probability of a draw from the term increases. But for large α, the data term will become prominent only when a substantial number of draws is already available, i. e. when it already constitutes a good approximation to G_0. For small α, the random measure will concentrate around the Dirac measures located at a small number of initial observations. This is again analogous to the finite-dimensional case, since the Dirac measures represent the extremal points of the infinite-dimensional probability simplex.

Dirichlet draws are not smooth. Let Θ be distributed according to a d-dimensional Dirichlet density $p(\theta|\beta,\pi)$. Assume that for some bin $i \in \{1,\ldots,d\}$, the expectation of the randomly assigned probability θ_i should be increased with respect to the current value. θ is normalized, and so the increase requires a decrease in expectation of values somewhere else. The values θ_j for individual bins in the Dirichlet distribution couple only through normalization. Such coupling does not single out any bin in particular, and on average, we have to expect every bin except θ_i to decrease. If one particular bin was more likely to decrease than others, the implication would be an additional interaction structure not given in Dirichlet distributions. In other words, the bins are anti-correlated. Indeed, the covariance of a Dirichlet variable is

$$\mathrm{Cov}\left[\Theta_i,\Theta_j\right] = -\frac{\pi_i\pi_j}{\beta+1} \qquad \text{for } i \neq j \; . \tag{2.4.10}$$

But this means in particular that neighboring bins have negative correlation, which is precisely the opposite of “smooth” behavior: To generate a smooth function, an increase at a given point should come with a simultaneous increase of its neighbors. This behavior is reflected in the limit, though it does not carry over precisely, but becomes more severe: A draw from the DP is discrete a. s. (Ferguson, 1973; Blackwell, 1973), i. e. representable as a countable sum of Diracs on the sample space Ω. Nonparametric Bayesian statisticians, motivated by the search for universal priors, tend to regard discreteness as a fundamental drawback. Somewhat ironically, discreteness is the property which makes the DP applicable to clustering problems. The interest of the machine learning community was raised by the very property that Bayesian statisticians have worked so hard to overcome.

2.4.5 References

Early Bayesian nonparametrics. The Dirichlet process (and the corresponding approach to priors) were introduced by Ferguson (1973), who attributes the problem idea to David Blackwell, and the solution (the Dirichlet process) to his own discussions with James B. MacQueen. Almost simultaneously with Ferguson's paper, a number of works appeared in a burst, including the proof of DP discreteness by Blackwell (1973), the Pólya urn interpretation by Blackwell and MacQueen (1973) and the DP mixture model described by Ferguson's student Antoniak (1974). From there on, interest in the statistics community focused primarily on overcoming the discreteness property of the DP, with models such as the DP mixture model, tailfree processes (Doksum, 1974) and Pólya trees (Ferguson, 1974). NTR processes were introduced by Doksum (1974), and the idea was taken up by Ferguson and Phadia (1979) in the context of survival analysis. The latter had previously been considered in a Bayesian nonparametric context by Susarla and Ryzin (1976), who apply the Dirichlet process to right-censored data and obtain a Kaplan-Meier estimator in the limit $\alpha \to 0$. The name "Chinese restaurant process", which has caused havoc in machine learning conference proceedings, is due to L. E. Dubins and J. Pitman. They developed the model, independent of the Dirichlet process, in a combinatorial manner somewhat resembling the infinite Pólya urn of Blackwell and MacQueen (1973). Their development is mentioned first by Aldous (1985), and later by Pitman (1995).

The extension theorem. The Kolmogorov theorem was proven originally for direct products of Euclidean axes (Kolmogorov, 1950). It was generalized consecutively to direct products of σ-compact measure spaces, and to complete separable metric spaces (Parthasarathy, 1967; Yamasaki, 1985). The theorem soon became the principal tool of construction in the theory of stochastic processes, though it has been partially superseded in this regard by stochastic differential equations (Gikhman and Skorohod, 1974; Øksendal, 1992). As mentioned above, the completely independent case of infinite product measures requires no assumptions on the topology of the underlying spaces. Between the case of fully independent variables on the one hand, and the Kolmogorov case of arbitrary dependence structure on the other hand, there are some intermediate results. For example, if the variables have a total order, each variable n can be conditioned on the previous $n-1$ variables. The projective family of Th. 23 is then replaced by a recursively defined sequence of random variables, specified by conditional distributions. Such variables have a projective limit theorem, with no regularity assumptions on the underlying spaces required (Ionescu Tulcea, 1950).

De Finetti's theorem and related symmetry results. De Finetti (1931) proves his theorem for binary random variables. Hewitt and Savage (1955) generalize to compact Hausdorff spaces, and show that the mixing distribution is uniquely determined. They also provide a rigorous formulation for the intuitive interpretation of the mixture as a convex combination, showing that the set of exchangeable measures is an infinite-dimensional convex polytope within the probability simplex. The extremal points of the polytope are the product measures. If exchangeability of individual elements of the sequence is substituted by a block-wise structure, called *partial exchangeability*, a similar result holds, with the product model replaced by Markov chains, which account for the block correlations (Diaconis and Freedman, 1980, 1984). An analogous theorem is available in continuous time (i. e. for sequences of random variables with index set $\mathbb{R}_+$): The continuous-time stochastic process analogue of a product distribution is a Lévy process. And sure enough, continuous-time exchangeable increment processes are mixtures of Lévy processes (Bühlmann, 1960). Exchangeable observations inherit a number of key properties from the independent case. In particular, they have a strong law of large numbers (Kingman, 1978) and a central limit theorem (Bühlmann, 1960). Other types of symmetries can have similar consequences. Ryll-Nardzewski (1957) proves a version of de Finetti's theorem establishing equivalence between conditional independence and contractability (the form of the de Fintteti theorem given in Th. 48). Kallenberg (2005) systematically studies invariance of random sequences under contractability, exchangeability and rotatability, and their mutual relations.

Bayesian nonparametrics in machine learning. In the machine learning context, Bayesian nonparametrics refers, first and foremost, to Dirichlet process mixtures. In machine learning, DP mixtures became popular only recently (Blei and Jordan, 2004; McAuliffe *et al.*, 2006), after Dirichlet distributions were considered as priors for multinomial topic models in text processing Zaragoza *et al.* (2003); Blei *et al.* (2003). The idea of infinite mixtures had previously been proposed in Bayesian machine learning, notably by Neal (1991) and Rasmussen (2000). Numerous constructions of new models have followed. At first, these were mostly hierarchical combinations of Bayesian nonparametric models (Teh *et al.*, 2004; Sudderth *et al.*, 2006) or modifications of DP mixtures in analogy to finite mixtures (Beal *et al.*, 2002; Orbanz and Buhmann, 2006). A readable and well-illustrated introduction to DPM models in machine learning is given by Sudderth (2006). Y. W. Teh has considered machine learning applications of models related to the DP that are available in the machine learning literature, in particular the coalescent of J. F. C. Kingman and the Pitman-Yor process (Teh, 2006; Teh *et al.*, 2008a). The first authors to suggest constructions of new models, rather than modifications of existing ones, are Neal (2003) and Griffiths and Ghahramani (2005). Neal (2003) defines a clustering model with a tree structure, generated by Brownian motions endowed with a splitting rule. Griffiths and Ghahramani (2005) generate infinite binary matrices at random by means of what is essentially a Lévy point process, shown to be conjugate to the beta process of Hjort (1990) by Thibaux and Jordan (2006).

Quantitative results. Aside from a large number of DP-based models and studies on algorithmic inference (cf. 2.6), the basic properties of DP models are still a subject of active research. A classic result of Diaconis and Freedman (1986) shows that a simple location estimate by means of a DP prior can be inconsistent. Even in the asymptotic limit of infinitely many observations, the data can be overruled by the prior assumption. The effect still seems far from being thoroughly understood. The original article is complemented by a spirited contributed discussion as to what causes the inconsistency, and by an author's rejoinder distinguished by its timeless style ("Well, *Krasker-Pratt*, lots of luck!"). Several discussants blame the DP's discreteness. The authors disagree in the rejoinder. Almost twenty years on, Ghosh and Ramamoorthi (2002) write in their monograph on the matter that, as a sufficient condition for consistency, they "believe that Diaconis and Freedman are correct in thinking that existence of density for random P is not enough." A number of quantitative results on posterior convergence have become available in recent years. Posterior convergence rates are addressed by Shen and Wasserman (2001) and Ghosal *et al.* (2002). They show that the convergence rate of the posterior, i. e. the rate with respect to sample size at which the posterior concentrates around the true model, is completely determined by the complexity of the nonparametric model and the prior probability of the true solution. Model complexity cannot be quantified, in the infinite-dimensional case, by counting dimensions. Instead, it is formalized in terms of metric entropy rates. The prior enters in so far as it has to place sufficient probability mass on a neighborhood of the true model. The consequence for Bayesian model specification is that, because the true solution is not known beforehand, the prior has to be chosen to put sufficient mass on the respective neighborhood of all models that constitute possible solutions. This notion, formalized as a prior concentration rate, quantifies the qualitative ideas discussed by Draper (1999). Kleijn and van der Vaart (2006) study the misspecification problem, i. e. the effect of the true model not being in the domain of the prior. The posterior is shown to concentrate, at a quantifiable rate, in the region of parameter space which is closest to the true model in a Kullback-Leibler sense. For other nonparametric Bayesian models, in particular those of Gaussian type, consistency results are available. These are often based on modifications of techniques already proven successful in classical nonparametrics. For example, Barron *et al.* (1999) apply Grenander sieves and entropy bracketing, previously applied for similar problems in the classical setting e. g. by van de Geer (1993). Results for Gaussian process regression are given by Diaconis and Freedman (1998); Choi and Schervish (2007).

Exponential family approximations and orthogonal expansions. An alternative approach for the definition of probabilities on measures is based on orthogonal expansions, by defining a set of orthogonal basis functions and estimating an expansion from the sample population. The number of basis functions used in the expansion can be based on the sample size, to provide better resolution for larger samples. Density expansions have been considered by Hall (1986), and later rose to prominence with the application of wavelet bases (Donoho *et al.*, 1996). A drawback is that linear basis approximations of densities can be negative. But long before wavelets were even introduced, a series of papers by Crain (1973, 1974, 1976a,b) studied expansion of the *logarithm* of the density in terms of orthogonal polynomials, also in a sample-size adaptive fashion. This is closely related to a basis expansion of the energy E in Eq. 2.2.9 (though direct expansion of the model logarithm targets the log-partition function along with the energy). It was already noted in Sec. 2.2.3 that exponential family models approximate the energy within a finite-dimensional linear subspace. Putting the different approaches together, one may consider an exponential family approximation with a sample-size adaptive expansion of the energy by means of orthogonal basis functions. That is, the sufficient statistic components $s_1(x), \ldots, s_d(x)$ are chosen e. g. as orthogonal polynomials. The actual dimension d grows with the sample size. In the limit of infinitely many observations, $d \to \infty$, but since the s_i form an orthonormal system in a suitable function space (L_2 or Sobolev), the scalar product converges, and the model is well-defined if the partition function integral converges (which need not be a trivial matter). Note that the model, when used in a Bayesian regime, satisfies Def. 20 of a nonparametric Bayesian model, provided that the sample-size adaptation rule is chosen such that the number of additional basis elements per additional observation is constantly bounded. Another Bayesian nonparametric approach using exponential families of adaptive dimension is the sieve prior method of Zhao (2000), which defines a prior on the overall model class by placing a prior on the model dimension.

Integral kernel push-forwards. Pillai *et al.* (2007) consider a Mercer kernel approach to the definition of nonparametric Bayesian priors on function spaces. The reproducing kernel Hilbert space of an arbitrary given kernel is shown to be spanned by the image of the set of signed Borel measures under the integral operator defined by the kernel. Available nonparametric Bayesian priors on the Borel measures then define image priors on the reproducing kernel Hilbert space.

Classical infinite-dimensional estimation. The problem of classical statistical estimation in infinite-dimensional spaces was first studied in depth by Grenander (1981). He considers in particular the properties of maximum likelihood estimation for infinite-dimensional parameters. The monograph provides a variety of results on estimation with Gaussian processes, and on general maximum likelihood estimation in the most common infinite-dimensional spaces (including ℓ_1 and ℓ_2, L_1 and L_2, BV, and the set of bounded linear operators). A common theme is that maximum likelihood estimators on infinite-dimensional spaces share the essential properties of their finite-dimensional counterparts, but that the proof techniques applicable in the finite-dimensional case are not adequate. To prove consistency properties of the estimators, Grenander (1981) develops his famous *method of sieves.* The proofs have to be adapted for each type of infinite-dimensional space in turn.

2.5 Mixture Models

Let $f_{X|z}$ be a parametric family of densities as in Def. 1, with observation variable X and parameter variable Z. If a distribution μ_{Z} for the parameter variable Z is given, the parameter may be integrated out of the model to obtain an unconditional distribution:

$$p(x) := \int_{\Omega_z} f_{X|z}(x|z) d\mu_{\mathrm{Z}}(z) \tag{2.5.1}$$

In common parlance, $f_{X|z}$ is *mixed against* μ_{Z}, and the model (2.5.1) is called a *mixture model* with *mixing distribution* μ_{Z}. The mixture model describes a two-stage process of data generation: Draw a parameter value z at random according to μ_{Z}, then generate an observation x according to the density $f_{X|z}(x|z)$.

The most commonly used types of mixtures are arguably finite mixture models (see below), for which the mixing distribution is discrete. An interesting example of a continuous mixture is Student's t-distribution, often advocated as a robust substitute for Gaussian priors by merit of its heavy tails. A t-model can be defined as a mixture, where the heavy tails are obtained by averaging over Gaussians of different variance. The example is included here because of its relevance for Bayesian modeling. Student distributions are not exponential family models, and do not have a conjugate prior. Nonetheless, they may be used as conjugate priors for Gaussian location parameters, because they are representable as mixtures of Gaussians, and mixture of conjugate priors are conjugate priors, resulting in a mixture posterior.

Example 26 (Continuous scalar mixtures as quotient models)**.** For Z with values in $\mathbb{R}_+$, (2.5.1) can be regarded as a quotient model. If X is a real-valued random variable with (differentiable) density $f(x)$, then for any positive scalar z, the scaled variable X/z has density $f(xz)z$. Hence if z is regarded as random, with distribution μ_{Z}, the random variable X/Z has density (2.5.1) with $f_{X|z}(x|z) := f(xz)z$. A particular case is Student's t-distribution, which can be obtained by mixing a normal against a gamma distribution on the variance parameter. A t-model describes the distribution of the t-statistic,

$$T := \frac{\sqrt{n}X}{\sum_{i=1}^{n} Y_i^2} \ , \tag{2.5.2}$$

where X, Y_i are independent, X is normal $\mathcal{N}(\mu, \sigma_X)$, and the Y_i are normal $\mathcal{N}(0, \sigma_y)$. The actual mixing variable is the sum in the quotient, $Z := \sum Y_i^2$, which is gamma distributed as $\mathcal{G}(n/2, 1/2\sigma_Y^2)$, and χ^2 in the special case $\sigma_y = 1$. Denote the densities of X and Z as $f(x|\mu, \sigma_x)$ and $m(z|n/2, 1/2\sigma_y^2)$, respectively. Then T has density

$$p(T = x|\mu, n/2, \sigma_y) := \int f(x|\mu, z) m(z|n/2, 1/2\sigma_y^2) dz \ . \tag{2.5.3}$$

The distribution given by $p(x|\mu, n/2, \sigma_y)$, or more generally by $p(x|\mu, \alpha, \beta)$ with $\alpha, \beta > 0$, is Student's t-distribution.

2.5.1 Finite Mixtures

Mixture models are called *finite* if the mixing distribution is categorical. That is, the parameter Z assumes only a finite number K of different values θ_k^*, each with fixed probability c_k. The density of μ_{Z} is then of the form

$$m_Z(z|c, \theta^*) = \sum_{k=1}^{K} c_k \delta_{\theta_k^*}(z) \ , \tag{2.5.4}$$

and the resulting mixture model (2.5.1) is

$$p(x|c, \theta^*) = \int F_{X|z}(x|z) m_Z(z|c, \theta^*) dz = \sum_{k=1}^{K} c_k f_{X|z}(x|\theta_k^*) \ . \tag{2.5.5}$$

A finite mixture model describes, by means of the sum, an exclusive conjunction of K "classes", distributed individually according to $f(x|\theta_k^*)$. The distribution of the class index k of x is multinomial in the mixture weights c_k.

2.5.2 Bayesian Mixtures

A mixture model becomes a Bayesian model if the mixing distribution is generated at random. For a finite mixture with a fixed number K of classes, random generation μ_{Z} amounts to random generation of the parameters θ^*_k and c_k in (2.5.4) above. The model commonly referred to as a *Bayesian mixture* in the literature is obtained from a finite mixture model by placing the respective conjugate prior on each parameter occurring in the model. The existence of a conjugate prior is guaranteed if $f_{X|z}$ is chosen as an exponential family model. The parameters of the model are then class parameters θ^*_k, and the mixture weights c_k. The conjugate prior for θ^*_k is determined by choice of $f_{X|z}$. Since the distribution of the class in a finite mixture is multinomial, the conjugate prior for the mixture weights c_k is a Dirichlet distribution. If the density of conjugate prior for θ^*_k is denoted $g(\,.\,|\lambda_k, y_k)$ and that of the Dirichlet by $g_{\mathrm{Dir}}(\,.\,|\beta, \pi)$, the Bayesian mixture posterior is, in full detail,

$$p(c_1, \dots, c_K, \theta^*_1, \dots, \theta^*_K | x_1, \dots, x_n, \beta, \pi, \lambda_1, \dots, \lambda_K, y_1, \dots, y_K) \propto \\ \Big(\prod_{i=1}^{n} \sum_{k=1}^{K} c_k f_{X|z}(x_i|\theta^*_k)\Big) \Big(\prod_{k=1}^{K} g(\theta^*_k|\lambda_k, y_k)\Big) g_{\mathrm{Dir}}(c|\beta, \pi)\ . \quad (2.5.6)$$

The priors g are *not* the mixing density m, which in this representation has already been integrated out, such that f is parameterized by θ^* rather than Z.

2.5.3 Dirichlet Process Mixtures

The model obtained by replacing the Dirichlet prior on the mixture weights in a Bayesian mixture model by a Dirichlet process is called a *Dirichlet process mixture*. Formally, the substitution is achieved by drawing the complete mixing distribution μ_{Z} in (2.5.1) from a Dirichlet process.

In contrast to the Bayesian mixture model above, in which the Dirichlet distribution explicitly generates the cluster proportions c_k, a sample from the Dirichlet process actually consists only of parameter values θ_i. The weights are generated implicitly. The clustering property of the DP aggregates observations into $K \leq n$ groups of identical values:

$$\sum_{i=1}^{n} \delta_{\theta_i} = \sum_{k=1}^{K} n_k \delta_{\theta^*_k}\ . \quad (2.5.7)$$

The cluster proportions c_k are computed as $c_k = \frac{n_k}{n}$. Since a single draw $z \sim \mathrm{DP}\,(\alpha G_0)$ from a Dirichlet process is generated as $z \sim G_0$, the expected measure G_0 assumes the role of the conjugate priors $g(\,.\,|\lambda_k, y_k)$ in the Bayesian mixture. If the density of G_0 is $g_0(\,.\,) := g(\,.\,|\lambda, y)$, this implies $\lambda_k = \lambda$ and $y_k = y$ for all k. That is, the hyperparameters cannot be specified individually for each class, as is possible for the Bayesian mixture model. The DP posterior has expected measure $\frac{1}{n+\alpha}(\sum \delta_{\theta_i} + \alpha G_0)$. Since the next sample is generated according to the expected measure, the mixing distribution μ_{z} has density

$$g(z|n_k, \theta_k^*, \alpha) = \frac{1}{\alpha + n}\Big(\sum_{k=1}^{K} n_k \delta_{\theta_k^*}(z) + \alpha g_0(z)\Big)\,. \tag{2.5.8}$$

The sum component is equivalent to the mixing density (2.5.4) of the finite mixture. A draw z according to g is drawn from the sum component with probability $\frac{n}{\alpha+n}$, in which case it will take one of the predetermined values θ_k^*. With probability $\frac{\alpha}{\alpha+n}$, the sample z is generated from G_0, and $z \neq \theta_k^*$ for all k unless G_0 is finite.

Remark 27 (Notation: Θ versus Z)**.** Parametric models in previous sections were denoted $\mu(X|\Theta)$, with X denoting observations and Θ a parameter variable. When Θ was itself parameterized, the hyperparameter was denoted Y. Mixtures add another layer of random values between X and Θ, in form of the mixing variable Z. Distinguishing between Z and Θ can be difficult, in particular in finite, Bayesian and Dirichlet process mixtures, for two reasons:

1. Z effectively takes values in the range of Θ.
2. In finite mixtures, Z can equivalently be assumed to take index values $1, \ldots, K$. This notation is also used in Dirichlet process mixture sampling algorithms.

Concerning (1). The mixing density in (2.5.4) is parameterized by Θ. A random draw $(\theta_1^*, \ldots, \theta_K^*, c_1, \ldots, c_K)$ determines the set of values θ_k^* which Z may assume, and their respective probabilities c_k. When the parametric prior G on Θ is replaced by a Dirichlet process, the finite set $\{\theta_1^*, \ldots, \theta_K^*\}$ is replaced by an infinite set $\{\theta_1^*, \theta_2^*, \ldots\}$. The set of weights $c_1, \ldots, c_K$, positive and normalized in the finite case, is replaced by the random measure drawn from the Dirichlet. The Dirichlet base measure, when regarded as a Bayesian parameter (and hence a random value itself), is an instance of Y. For both Bayesian and DP mixtures, the mixing measure is determined as a random draw from Θ. The domain Ω_θ of Θ formally decomposes as $\Omega_\theta = (\Omega_\theta^*)^K \times \mathrm{Sim}(\mathbb{R}, K)$. Here, Ω_θ^* is the range of θ_k^*, such as $\mathbb{R}$ or a vector space, and the simplex $\mathrm{Sim}(\mathbb{R}, K)$ accounts for the weight vector c. In the

DP limit, the pair is replaced by the set of measures $\mathcal{M}^1_+(\Omega^*_\theta)$, such that Ω^*_θ still specifies the range of θ^*_k. In both cases, the range of Z is then a subset of the domain Ω_θ of Θ. In the finite mixture case, the subset is known explicitly. In the Dirichlet case, only a finite subset can be determined by drawing a sample of size n from the Dirichlet via its base measure. The Sethuraman (1994) representation casts a Dirichlet random draw in a form roughly corresponding to (2.5.4). The fact that Z takes values θ^*_k can make notation hard to follow, since random values x are generated according to mixture components of the form $f(x|\theta^*_k)$, where the abstract density is $f(x|z)$.

Concerning (2). In a finite or Bayesian mixture, with $\{\theta^*_1, \dots, \theta^*_K\}$ given explicitly, Z is often assumed to take values in $1, \dots, K$, formalizing the notion of a cluster assignment or indicator variable. This is particularly useful in EM algorithms or blocked Gibbs samplers (Sec. 2.6), where such indicator variables are generated explicitly. For a Dirichlet process mixture, it is more difficult, but still used in sampling algorithms, where an estimate of the set $\{\theta^*_1, \theta^*_2, \dots\}$ is explicitly generated and finite due to finite sample size.

2.5.4 References

The concept of parametric mixture models dates back, at least, to the work of Newcomb (1886).[7] A few years later, Pearson (1894) considered estimation of parameters in a two-component mixture. His work predates Fisher's work on the maximum likelihood method, and he proposes a moment-based fitting procedure. Moment-based fitting is not typically used in the mixture model literature anymore, but discussed for example by Titterington *et al.* (1995). Though a lion share of work on mixture models has been devoted to inference algorithms (see below), the statistical properties of the models and the model order selection problem have received considerable attention. The relevant literature, in particular on the Gaussian mixture, is too large to even be sketched on a few pages, and finds application throughout statistics, machine learning, speech and signal processing, data mining, and so forth. See for example McLachlan and Peel (2000) for an overview.

Bayesian and DP mixtures. Though the Dirichlet process and the DP mixture model were introduced well over thirty years ago (Ferguson, 1973; Antoniak, 1974), DP mixtures received only limited attention until the early

[7] To each estimation error, Newcomb relates a quantity which he calls the *evil attached to an error*, and his article consequently contains paragraphs entitled "Algebraic Expression for the Evil" and "Approximation Expression for the Evil". He attributes the concept of the evil to Gauss, who introduces it in his treatise "Theoria Combinationis Observationum, etc., Pars prior" under the name *jactua* – loss.

1990s. Interest in DP mixtures then increased suddenly, as feasible inference algorithms became available in the wake of the Gibbs sampler (MacEachern, 1994; Escobar, 1994; Escobar and West, 1995). An infinite mixture model which is (algorithmically) equivalent to the DP mixture was actually introduced by Neal (1991), who reported later to have been unaware at the time of the DP (Neal, 2000). At about the same time as the first DPM Gibbs samplers, the Bayesian mixture model appeared in a somewhat different part of the Bayesian community (Lavine and West, 1992; Mengersen and Robert, 1995).

Text processing: pLSA and LDA. The finite mixture of multinomial distributions, with a standard EM algorithm for inference, triggered an avalanche of publications in natural language processing when Hofmann (1999) published it under the name *probabilistic latent semantic analysis* (pLSA). The state-of-the-art approach to "semantic" modeling of text documents at the time, called *latent semantic analysis* (Deerwester *et al.*, 1990), represents a text document by its word occurrence histogram. The histograms are regarded as vectors in a high-dimensional vector space (each axes represents a word in the vocabulary). Similarity between documents is measured by means of a scalar product. Latent semantic indexing performs a dimension reduction by computing a singular value decomposition and retaining only a number of large singular values, which would amount to principal component analysis if all matrices involved were symmetric. The dimension reduction projects onto a subspace, each axis of which is a linear combination of the original axes. The original axes each represent occurrence of a single word, and the new axes therefore weighted combinations of words. These are interpreted as "topics" in the latent semantic analysis model. Hofmann (1999) notes that the effective domain of the histograms is a scaled simplex. He substitutes convex for linear combinations, expectation-maximization for singular value decomposition and Kullback-Leibler divergence for Euclidean distances, to obtain a mixture model in which each multinomial distribution represents a topic. The parameter vector of a multinomial component is interpreted as a finite probability distribution over words, characterizing the topic by how likely a given word is to occur in an associated document. The Bayesian mixture extension of the pLSA multinomial mixture, with the priors $g(\,.\,|\lambda_k, y_k)$ on the component parameters θ_k^* all chosen as uniform distributions, was published as a model for text processing by Blei *et al.* (2003) under the name *latent Dirichlet allocation* (LDA).

2.6 Latent Variable Algorithms

Latent variable algorithms are the methods of choice for algorithmic inference in mixture models. They include certain Gibbs samplers, the EM family of algorithms, and many "variational" methods popular in machine learning. Suppose the mixing distribution in (2.5.1) is parameterized as $\mu_{Z|\Theta}(z|\Theta = \theta)$, with conditional density $g(z|\theta)$ w. r. t. Lebesgue measure:

$$p(x|\theta) = \int_{\Omega_z} f_{X|Z}(x|z) d\mu_{Z|\Theta}(z|\Theta = \theta) = \int_{\Omega_z} f_{X|Z}(x|z) g_{Z|\Theta}(z|\theta) dz \tag{2.6.1}$$

The objective of a mixture model estimation algorithm is to determine the parameters θ of the mixing distribution from observations $x_1, \ldots, x_n$.

2.6.1 General Strategy

The precise objective of parameter estimation may vary, depending on whether estimation is approached in terms of maximum likelihood, posterior approximation or posterior maximization. The common strategy of most algorithms (and in particular of all algorithms considered in this thesis) is a *latent variable* approach. The name refers to the mixing variable z, considered "latent" because it is neither an estimation target, nor determined explicitly by observation. Latent variable methods approximate the integral in (2.6.1), or its θ-derivative, by substituting some point estimate of z_0 in the integrand. To take the variance w. r. t. to z into account, procedures may average over multiple estimates z_0. The essential ingredients of latent variable methods are coordinate or block optimization strategies, and – depending on the approach – majorization or random sampling techniques.

Coordinate Relaxation and Blocking

The core technique of latent variable methods is the *coordinate* or *block optimization* approach, which ultimately dates back to Gauss and the Gauss-Seidel solver. The objective is to optimize some function $A : \Omega^n \to \mathbb{R}$, that is, to determine

$$\hat{x} := \arg \min_{x \in \Omega^n} A(x_1, \ldots, x_n) \; . \tag{2.6.2}$$

A *coordinate optimization* strategy approaches the problem by an iterative procedure. In each iteration step t, the method cycles over the coordinates $i = 1, \ldots, n$ and computes

$$x_i^t = \arg \min_{x_i \in \Omega} A(x_1^t, \ldots, x_{i-1}^t, x_i, x_{i+1}^{t-1}, \ldots, x_n^{t-1}) \; . \tag{2.6.3}$$

That is, the function is optimized along each axis separately, by substituting estimates of the elements of x for all other entries. Each new estimate is immediately taken into account (hence the $(i-1)$ leading arguments of A are indexed by t, the the trailing ones by $(t-1)$). The Gauss-Seidel solver for linear equation systems is a special case, with A linear. Coordinate optimization methods are also referred to as "coordinate relaxation", because the original problem of jointly optimizing (2.6.2) is "relaxed" to yield the presumably less difficult problems (2.6.3).

Coordinate optimization is straightforwardly generalized along several lines: Block optimization, descend methods, and sampling strategies. The coordinate-wise optimization may be replaced by *block optimization* (or *block relaxation*). The n axes are grouped into $K \leq n$ blocks of summary coordinates $\tilde{x}_1, \dots, \tilde{x}_K$, corresponding to a subdivision of the space according to $\Omega^n = \Omega^{n_1} \times \cdots \times \Omega^{n_K}$. Optimization of $A(x) = A(\tilde{x}_1, \dots, \tilde{x}_K)$ is then performed block-wise,

$$\tilde{x}_k^t = \arg \min_{\tilde{x}_k \in \Omega^{n_k}} A(\tilde{x}_1^t, \dots, \tilde{x}_{k-1}^t, \tilde{x}_k, \tilde{x}_{k+1}^{t-1}, \dots, \tilde{x}_K^{t-1}) \,. \tag{2.6.4}$$

Apparently, coordinate methods are included as the special case $K = n$. *Descend algorithms*, as opposed to optimization algorithms, are methods which compute updates $\tilde{x}_k^t$ such that A decreases, but is not necessarily minimized. *Coordinate samplers* and *block samplers* draw the updates $\tilde{x}_k^t$ at random, from a distribution chosen such that the result decreases A in probability.

Latent variable methods use two blocks, representing the parameters θ and the latent variables z. EM-type algorithm apply block *maximization* to a function A chosen to minorize the objective function. The Gibbs sampler is a coordinate descend algorithm with stochastic updates. The blocked Gibbs sampler combines both approaches, using a latent variable approach with two blocks. Within the blocks, stochastic coordinate updates are used for the latent variables, and stochastic block updates for the parameters.

Majorization and Augmentation

If a minimization target function A is hard to optimize directly, many optimization strategies replace A with a simpler approximation B, then optimize B instead of A. For purposes of optimization, approximations are best chosen as upper bounds (for minimization). If $B(x) > A(x)$ for all x, then even if the approximation quality at an estimated optimizer x^* cannot be quantified, the method always guarantees the value $A(x^*)$ to be at most $B(x^*)$. Such upper bounds can be constructed by means of inequalities. In statistics, most constructions are of course based on Jensen's inequality, but other examples are Cauchy-Schwartz, Bernstein (for maxima of polynomials) or Hölder.

For latent variable methods (and many others), the majorizing function is constructed on a larger space, i. e. $A(x)$ is upper-bounded by a function $B(x, y)$. Such techniques are known as *data augmentation* in the statistics literature. The name emphasizes a perspective which regards the second argument y of B as an artificial addition to the "true" variable x, but for application to mixture models, the second argument is naturally provided by the mixing variable. Augmentation methods require the upper bound to be tight, that is, a function B is called a majorizing function for A if

$$A(x) = \min_y B(x, y) \ . \tag{2.6.5}$$

Then $\min_x A(x) = \min_{x,y} B(x, y)$, and minimization is performed by means of the block optimization (2.6.3) with two blocks representing x and y. Majorizing functions constructed by means of an inequality often satisfy (2.6.5), since many standard inequalities become equalities in well-defined special cases. Consider Hölder's inequality, for example: If A, B are functionals and the arguments x, y are normalized l_p functions, then equality in Hölder holds if and only if $x = y$. The minimum (2.6.5) would then take the form $\min_y B(x, y) = B(x, x)$. This is a stronger form of (2.6.5). The minimum is assumed for $y = x$. The stronger form is often encountered as an additional assumption for augmentation methods, particularly for EM and related algorithms. The requirements on B are then

$$A(x) = B(x, x) \qquad \text{and} \qquad A(x) \leq B(x, y) \ . \tag{2.6.6}$$

This case is worth noting, because it turns block-wise minimization into a fixed point iteration. If, during step t of the iteration,

$$x^t := \arg\min_x B(x, y^{t-1}) \ , \tag{2.6.7}$$

then

$$A(x^t) = B(x^t, x^t) = \min_y B(x^t, y) \ . \tag{2.6.8}$$

The two-step block optimization reduces to a single step

$$y^{t+1} := \arg\min_x B(x, y^t) \ . \tag{2.6.9}$$

This is an application of the Newton fixed point iteration (given by $x^{t+1} := f(x^t)$ for an arbitrary function f) to the function $y \mapsto \arg\min_x B(x, y)$.

2.6.2 The EM Algorithm

EM algorithms optimize (maximize) functions of the form

$$f(\theta) = \int_{\Omega_z} h(\theta, z) dz \ . \tag{2.6.10}$$

In statistical applications, this is of course a mixture as in (2.6.1). For convenience, denote in the following all density functions by p, in particular, $p(x|z) = f_{X|Z}(x|z)$ and $p(z|\theta) = g_{Z|\Theta}(z|\theta)$. Then in the integral,

$$h(\theta, z) = p(x, z|\theta) = p(x|z)p(z|\theta) \ , \tag{2.6.11}$$

and consequently $f(\theta) = p(x|\theta)$. However, the algorithm does not rely on the special properties of densities. The only requirements are $h(\theta, z) \geq 0$ and $f(\theta) < \infty$, and the principles underlying the algorithm can be abstracted from statistical applications.

The integral f is optimized by optimizing its logarithm $l(\theta) := \log f(\theta)$. For any two values θ and θ',

$$l(\theta) - l(\theta') = \log\Big(\frac{\int \frac{h(z,\theta')}{h(z,\theta')} h(z,\theta) dz}{\int h(z,\theta')dz}\Big) \overset{\text{Jensen}}{\geq} \frac{\int h(z,\theta') \log\big(\frac{h(z,\theta)}{h(z,\theta')}\big) dz}{\int h(z,\theta')dz} \ . \tag{2.6.12}$$

Equivalently,

$$l(\theta) \geq l(\theta') - \frac{\int h(z,\theta') \log h(z,\theta') dz}{\int h(z,\theta')dz} + \frac{\int h(z,\theta') \log h(z,\theta) dz}{\int h(z,\theta')dz} =: B(\theta, \theta') \ . \tag{2.6.13}$$

Equality holds if (and only if) $\theta = \theta'$, hence $l(\theta) = \max B(\theta, \theta')$, and B satisfies a condition of the form (2.6.6), with minimization replaced by maximization. In the particular case of mixture model estimation, $h(z, \theta) = p(x, z|\theta)$. The only component of $B(\theta, \theta')$ depending on θ, i. e. the final quotient in (2.6.13), can be rewritten as

$$\frac{\int h(z,\theta') \log h(z,\theta) dz}{\int h(z,\theta')dz} = \mathbb{E}_{z|x,\theta'} \left[\log p(x, z|\theta)\right] \ . \tag{2.6.14}$$

The conditional expectation is typically denoted $\mathbb{E}_{z|x,\theta'} \left[\log p(x, z|\theta)\right] =: Q(\theta, \theta')$ in the EM literature. The classic EM algorithm of Dempster *et al.* (1977) iterates

1. (E-step) Construct $Q(\theta, \theta^t) = \mathbb{E}_{z|x,\theta^t} \left[\log p(x, z|\theta)\right]$ as a function of θ.

2. (M-step) Compute $\theta^{t+1} := \arg\max_\theta Q(\theta, \theta^t)$.

Maximizing $Q(\theta, \theta^t)$ maximizes $B(\theta, \theta^t)$, such that $l(\theta^t) = B(\theta^t, \theta^t)$ for any θ^t computed by the M-step. EM convergence results show that B, and hence l, never decreases under iteration.

Estimating a Finite Mixture

The most prominent application of the EM algorithm is approximative maximum likelihood estimation for the parameters $(c, \theta^*) =: \theta$ of a finite mixture model, as defined in Sec. 2.5.1. The density of the model is

$$p(x|c, \theta^*) = \sum_{k=1}^{K} c_k f(x|\theta_k^*) \ . \tag{2.6.15}$$

This form corresponds to the density p in 2.5.1, with the variable Z already integrated out of the model. Application of the equations above requires an explicit representation of Z, to compute the expectation $\mathbb{E}_{z|x,\theta'}\left[\log p(x, z|\theta)\right]$. Formally, the random variable Z takes values in $\{\theta_1^*, \dots, \theta_K^*\}$, with its distribution determined multinomially by the weights c. It will be convenient to represent Z instead as a one-count histogram of dimension K, such that $z_i = (0, \dots, 0, 1, 0, \dots, 0)$ for observation x_i. Intuitively, z_i encodes the assignment to a component, with $z_{ik} = 1$ if and only if observation x_i is assigned to mixture component k, and therefore to parameter θ_k^*. Since the values of all θ_k^* are given, this is equivalent to a representation where Z takes values $z = \theta_k^*$. Whenever the binary value has to be converted to an index, we will write $k(z_i)$ for the index k with $z_{ik} = 1$. The joint density of x_i and z_i is rewritten as

$$p(x_i, z_i|\theta) = p(x_i|z_i)p(z_i|\theta) = p(x_i|\theta_{k(z_i)}^*)\underbrace{p(k(z_i)|c)}_{=c_{k(z_i)}} \ . \tag{2.6.16}$$

Due to the 0-1 structure of z_i, the density of k conditional on c and x_i (instead of just c above) is

$$p(k(z_i)|c) = \mathbb{E}_{z_i|x_i,\theta'}\left[z_{ik}\right] \ . \tag{2.6.17}$$

Then the conditional expectation $\mathbb{E}_{z_i|x_i,\theta'}\left[\log p(x_i, z_i|\theta)\right]$, for a single observation x_i, is

$$\begin{aligned}\mathbb{E}_{z_i|x_i,\theta'}\left[\log p(x_i, z_i|\theta)\right] &= \sum_{z_i} p(z_i|x_i,\theta)\log p(x_i, z_i|\theta)\\ &= \sum_{k=1}^{K} p(k(z_i)=k|x_i,\theta)\log(p(x_i|\theta_k^* p(k(z_i)=k|c))\\ &= \sum_{k=1}^{K}\mathbb{E}_{z_i|x_i,\theta'}\left[z_{ik}\right]\log p(x_i|\theta_k^*) + \sum_{k=1}^{K}\mathbb{E}_{z_i|x_i,\theta'}\left[z_{ik}\right]\log c_k\ .\end{aligned} \tag{2.6.18}$$

For multiple observations,

$$\mathbb{E}_{z|x,\theta'}\left[\log\prod_{i=1}^{n} p(x_i, z_i|\theta)\right] = \sum_{i=1}^{n}\mathbb{E}_{z|x,\theta'}\left[\log p(x_i, z_i|\theta)\right]\ . \tag{2.6.19}$$

EM is a Block Optimizer

The block optimization character of the EM algorithm may not be immediately apparent, because two instances θ and θ' of the same variable appear in the target function $Q(\theta, \theta')$, and because the E-step seems to compute an expectation rather than a maximizer. The block structure becomes more recognizable if Q is rewritten as a function of two distinct variables θ and ξ,

$$Q(\theta, \xi) = \mathbb{E}_{z|x,\xi}\left[\log p(x, z|\theta)\right]\ . \tag{2.6.20}$$

A cycle of the block optimization procedure works as follows:

1. θ-block: For some fixed value of ξ, compute $\theta^t := \arg\max_\theta Q(\theta, \xi)$ (M-step).
2. Substitute $\theta = \theta^t$ into Q. The substitution adjusts the majorizing function to the new iterate θ^t.
3. ξ-block: Compute $\arg\max_\xi Q(\theta^t, \xi)$. But we know that the maximum is attained for identical values of both arguments, and hence for $Q(\theta^t, \theta^t)$. So without additional computation, substitute $\xi := \theta^t$.

Steps (2) and (3) together comprise the E-step of the EM algorithm. After execution of (3), $Q(\theta, \xi) = Q(\theta, \theta')$. The state $Q(\theta, \theta')$ is the input for the following M-step, matching the standard EM notation. The block optimization is obfuscated by standard notation, because it emphasizes the computation of the expectation in the E-step, for two reasons: First, the statistical intuition of how the algorithm operates relies on the expectation.

Second, the EM method is actually a family of algorithms applicable to a variety of problems. The general formulation of the method according to Dempster *et al.* (1977) does not specify (i) how to compute the expected value and (ii) how to solve the maximization problem in the M-step. Solution for both problems have to be provided in order to adapt the method to a given problem. It seems natural to formulate the algorithm in a two-step manner such that each step corresponds to one of these open problems, as in the E/M-step representation. Formulating it as a two-step block optimizer means that one of the steps (maximization w. r. t. ξ) is trivial. Nonetheless, the formulation as a block optimizer shows how and why the algorithm works, and clarifies its relation to other methods.

2.6.3 Blocked Gibbs Samplers

The Gibbs sampler applies the coordinate-wise ascent approach to the probabilistic ascent of Markov chain samplers. The great merit of the Gibbs sampling approach is to turn the design of a general MCMC sampler from something of an art form into a question of computing the so-called full conditionals of the target distribution. Assume the target distribution (the distribution from which the algorithm is supposed to generate random samples) has density $p(x_1, \dots, x_n)$. The design of a Metropolis-Hastings sampler for p, for example, requires (i) guessing a proposal distribution and (ii) showing that the proposal distribution results in a sampler with stationary distribution p. Gelfand and Smith (1990) showed that a coordinate-wise approach that successively generates x^t as

$$x_i^t \sim p(x_i | x_1^t, \dots, x_{i-1}^t, x_{i-1}^{t-1}, \dots, x_n^{t-1}) \qquad (2.6.21)$$

always results in a sampler with stationary distribution p. The conditional distribution p of x_i conditional on the current state of all other coordinates is commonly referred to as the *full conditional* of x_i. The additional randomized acceptance step performed by the general Metropolis-Hastings sampler is not required, because substitution of the full conditionals into the acceptance probability formula always results in an acceptance probability of one. Apparently, the algorithm is useful only if the full conditionals can be both computed and sampled in a reasonably simple manner. The widespread use of the Gibbs sampler shows that this is indeed the case for a wide variety of models, which is, on second glance, not particularly surprising. In the extreme case of independent coordinates, when the model density is a product, this apparently is the case if the product components are simple. Such models could, of course, have been sampled without development of MCMC methods. But most models of dependent random variables model restricted rather than full interactions. In particular, this is the case for the two examples of interest in the following, mixture models and Markov random fields. In a mixture model, the full conditionals are formulated as distributions conditional on an estimate of the model parameters, and hence decouple by conditional independence. In Markov random fields, the full conditional of x_i degenerates to a distribution conditional on the Markov blanket of x_i, which typically consists of a small number of points.

Just like the deterministic coordinate optimization scheme (2.6.2), the coordinate-wise Gibbs sampler may be generalized to a blocked version. The convergence behavior of Gibbs samplers depends on the correlation between states of consecutive iterations (Schervish and Carlin, 1992; Goodman and Sokal, 1989). If the components collected in each block are strongly correlated, blocked updates can result in substantial gains in convergence rate (Liu *et al.*, 1994; Roberts and Sahu, 1997). Mixture models have a natural correlation structure, and samplers using block structures defined by the assignment of data to mixture components have emerged as the method of choice for most mixture inference problems. Formally, blocked Markov chain algorithms can be derived by defining a Markov chain operating without blocks, and identifying its transition kernel. The product σ-algebra in the joint observation space Ω_x^n (for n observations) is then coarsened to a suitable σ-subalgebra. This is done by defining a measurable partition of the space, which corresponds to the block structure, and choosing the σ-algebra generated by the partition. The Markov kernel of the original chain induces a unique Markov kernel on the coarsened σ-algebra. The blocked Markov chain sampler is the one induced by the induced, coarsened kernel (MacEachern, 1994).

Posterior estimation in Bayesian mixtures

Implementing a blocked Gibbs sampler is much simpler than the formal approach sketched above may seem to suggest, and does not explicitly rely on the definition of a transition kernel. The blocks are defined by mixture assignments. They generally vary in size as assignments change throughout iteration. Each iteration therefore has to consist of two distinct steps, one determining the mixture assignments defining the blocks, and one performing the actual block updates. For Bayesian mixtures, the block update means sampling from the posterior of the corresponding mixture component, under those observations currently assigned to the block.

Consider a Bayesian mixture model as defined in Sec. 2.5.2. The component densities are denoted $f(x|\theta_k^*)$, the mixture weights c_k, and the Dirichlet prior on the weights by $g_{\text{Dir}}(c|\beta,\pi)$. Each component $f(\,.\,|\theta_k^*)$ has a corresponding conjugate prior $g(\theta_k^*|\lambda, y_k)$. A blocked Gibbs sampler for this model estimates a joint posterior on the parameters c and θ^* by iterating the following steps:

1. *Assignment step.* For $i = 1, \dots, n$, compute
$$q_{ik} := \frac{c_k f(x_i|\theta_k^*)}{\sum_{l=1}^{K} c_l f(x_i|\theta_l^*)} \tag{2.6.22}$$
Sample
$$z_i \sim (q_{i1}, \dots, q_{iK}) \tag{2.6.23}$$
2. *Parameter update step.* Sample
$$\begin{aligned}\theta_k^* &\sim g(\theta_k^*|\lambda, y_k) \prod_{i|z_i=k} f(x_i|\theta_k^*) \qquad \text{for } k = 1, \dots, K \\ (c_1, \dots, c_k) &\sim g_{\text{Dir}}(\,.\,|\beta\pi_1 + n_1, \dots, \beta\pi_k + n_k)\end{aligned} \tag{2.6.24}$$

The actual Gibbs sampling step is the second one, computing parameter updates for each block given the block structure. The first step is an auxiliary step to determine the blocks. To derive it formally from the mixture model, the observation x has to be integrated out of the joint density $p(x, z|\theta)$, conditional on $\theta = (c, \theta^*)$, since estimates of θ_k^* are given by the current state of the chain. Since x_i is fixed by observation, $p(x, z|\theta)$ is integrated against δ_{x_i}, which results in a multinomial distribution parameterized by q as defined in (2.6.27).

It is interesting to note that the literature on Bayesian mixtures started to develop in earnest only after the work of Gelfand and Smith (1990) on the Gibbs sampler. Samplers offer a mode of inference that avoids expansion of the product in (2.5.6), the main obstacle to fully Bayesian inference of the model in practice. Robert (1995) points out that (2.5.6) decomposes into K^n terms
$$\prod_{i=1}^{K} c_k^{z_k + n_k - 1} g(\theta_k^*|\lambda_k + n_k, y_k + n_k \bar{x}_k) \ , \tag{2.6.25}$$
e. g. about one million terms for two mixture components and twenty observations. Each term represents one possible combination of assignments of the observations to mixture components. In the blocked Gibbs sampler, the complexity is reduced to a single term, by forcing each observation to "select" one component. A comparison of the Gibbs sampling algorithm above and the EM algorithm in Sec. 2.6.2 shows a striking similarity between the two methods: Substitution of the sampling steps (2.6.28) and (2.6.30) by

maximization steps yields a binary assignment variant of the maximum a posteriori EM algorithm (also called MAP-EM). On clusters of reasonable size (i. e. if a reasonable number of observations is assigned to each mixture component), the results are virtually identical. The number n_k of points in cluster k increases the concentration parameters $\lambda_k + n_k$ of the posterior components, such that the posterior of a large cluster peaks sharply around its mode. Maximization and sampling then yield approximately identical values up to negligible probability. If clusters are well-separated, maximization and sampling of component assignments is also nearly equivalent. In other words, the algorithms behave very similarly for problems with reasonably-sized, well-separated clusters, which are precisely those problems on which either algorithm works reliably. Hence even without availability of the Gibbs sampler, approximate estimation of Bayesian mixtures should arguably have been within reach since the 1970s. One possible explanation, though unsubstantiated by evidence, is that sentiments in the Bayesian community demanded a sampler, even if that sampler behaves much as an EM point estimation algorithm would behave.

Posterior estimation in Dirichlet process mixtures

A Gibbs sampler for a Dirichlet process mixture can be derived in perfect analogy to the Bayesian mixture above. The model differs from the finite Bayesian mixture in that the mixing distribution G is drawn from a Dirichlet process DP (aG_0), instead of being generated parametrically from a product prior on (θ^*, c). In the posterior expectation

$$g(z|n_k, \theta_k^*, \alpha) = \frac{1}{\alpha + n}\Big(\sum_{k=1}^{K} n_k \delta_{\theta_k^*}(z) + \alpha g_0(z)\Big) . \qquad (2.6.26)$$

of the DP mixing density, the algorithm can treat the sum term just like a Bayesian mixture, but has to be modified to account for the second term αg_0. A draw from the z first has to select one of the components $\delta_{\theta_k^*}$ or g_0 at random, which is a multinomial decision just like in a standard mixture model. The only required modification of the sampling step for z_i is to add an additional component to the multinomial distribution to represent g_0. The component is denoted q_{i0} below. The implied meaning of $z_i = k$ is that x_i is assumed to have been generated according to the model parameter θ_k^*, so z_i assigns a parameter to an observation. A multinomial sampling value $z_i = 0$ models a parameter value not yet observed, so the parameter has to be generated from the base measure G_0. The algorithmic result is the creation of a “new cluster”.

1. *Assignment step.* For $i = 1, \dots, n$, compute

$$\tilde{q}_{i0} := \alpha \int f(x_i|z) G_0(z) dz \qquad \text{and} \qquad \tilde{q}_{ik} := n_k f(x_i|\theta^*_k) \quad (2.6.27)$$

   Normalize $q := \frac{\tilde{q}}{|\tilde{q}|}$ and sample

$$z_i \sim (q_{i0}, \dots, q_{iK}) \quad (2.6.28)$$

   If $z_i = 0$, sample

$$\theta^*_{K+1} \sim G_0 \quad (2.6.29)$$

   and set $z_i := K + 1$ and $K := K + 1$.
2. *Parameter update step.* Sample

$$\theta^*_k \sim g(\theta^*_k|\lambda, y_k) \prod_{i|z_i=k} f(x_i|\theta^*_k) \qquad \text{for } k = 1, \dots, K \quad (2.6.30)$$

The way new clusters are created by the algorithm is in practice influenced significantly by the coordinate optimization strategy of immediately incorporating changes: Clusters created by a sampling result $z_i = 0$ are available as mixture components for z_{i+1}, and not aggregated over a complete sweep $i = 1, \dots, n$. In principle, a total of n new components may be created during a single sweep. Scanning the data in index order is justified by arguing that data drawn from a Dirichlet process mixture is exchangeable, and the stationary distribution of the Markov chain sampler is invariant under permutations of the data. In practice, random scans may still be helpful, because the order of the data from one iteration to the next and may avoid unnecessary trapping states.

2.6.4 Convergence

The generality of basic convergence results for block optimization can be surprising, considering the amount of literature dedicated, for example, to the convergence of Gibbs samplers. Consider the block minimizer in (2.6.4). Suppose the optimization starts with initial value x_0, and that the optimization target function $A(x)$ is continuous in x. If the set of points with images "below the level set" of x_0, i. e. the set of points $\{x|A(x) \leq A(x_0)\}$, is compact, then the sequence $A(x^t)$ of function values under iteration is guaranteed to converge (because compactness and continuity imply boundedness from below). The sequence x^t has an accumulation point (by Bolzano-Weierstrass), and the value of A at the accumulation point is identical to the limit of A (by continuity). See de Leeuw (1994). Less trivial results are described, for example, by Bezdek *et al.* (1987). Many algorithms use more general block structures, in particular blocks chosen *adaptively* during the iteration (such as the ones used by the mixture model Gibbs sampler). The convergence behavior of deterministic optimization algorithms of this type is studied by Fiorot and Huard (1979).

The literature on convergence of Markov chain Monte Carlo algorithms in general, and the Gibbs sampler in particular, is vast and growing; for an overview, see for example Liu (2001), or the beautiful review of Hobart and Jones (2001). For diagnostic sampling methods, which attempt to determine convergence at runtime, see Goodman and Sokal (1989) and Cowles and Carlin (1996). A key property of the Gibbs sampler is the dependence of its efficiency on the correlation between components (Schervish and Carlin, 1992). The mixture model Gibbs sampler and other blocking strategies attempt to exploit this by adaptively grouping highly correlated components (Liu, 1994; Liu *et al.*, 1994; Roberts and Sahu, 1997).

The role of convexity in latent variable methods tends to be slightly over-emphasized in the machine learning literature. Though often helpful to guarantee, for example, a well-behaved form of a majorizing function, convexity is not essential. The EM algorithm, for example, does exploit concavity of the logarithm (by Jensen's inequality) for majorization, but the required property here is the applicability of *some* inequality to construct a majorizing function. Similarly, "entropic" interpretations of the EM algorithm, such as the alternating Kullback-Leibler minimization of Csiszár and Tusnády (1984), are useful and appealing – but the EM algorithm remains applicable if the functions involved are not probability densities, in which case entropic interpretations are no longer valid. The particular importance of convexity in optimization is due to the fact that convex functions allow their global properties to be deduced from local ones, and none of the algorithms discussed here relies on this principle.

2.6.5 References

The EM was first identified as a generally applicable family of algorithms for maximum likelihood approximation in mixtures by Dempster *et al.* (1977), who presented it as a unified generalization of a number of more specialized algorithms available in different fields, in particular the Baum-Welch algorithm for hidden Markov models (Baum and Petrie, 1966; Baum *et al.*, 1970). Apart from the algorithm, the paper presented a proof of local convergence, which asserts that (i) the model likelihood never decreases between consecutive steps of the algorithm and (ii) that the algorithm always converges to a local maximum. Unfortunately, the latter result is incorrect, due to a misapplication of the triangle inequality in the proof. Wu (1983) shows that the result can be recovered under additional regularity assumptions on the augmented likelihood function. Examples of the EM algorithm converging to a saddle point, or even a local minimum of the likelihood, have been reported in the literature (Arslan *et al.* (1993); also reproduced in the well-known monograph by McLachlan and Krishnan (1997)). They should be taken with a grain of salt, since the example in question is an EM algorithm which performs maximization in its M-step by numerical optimization, instead of computing an analytically derived maximizer, such as the EM algorithm for the Gaussian mixture does. Convergence to a local minimum is achieved in (Arslan *et al.*, 1993) by carefully tuning the parameters such that the gradient descent algorithm executed in the M-step precisely hits the local minimum. The result exemplifies the stepsize problem in gradient descent, rather than an inherent property of the EM algorithm. The problem does not occur for M-steps evaluating a closed-form maximizer.

Bayesian inference in mixtures generally builds on the work of Tanner and Wong (1987); Swendsen and Wang (1987) and, in particular, of Gelfand and Smith (1990). For general overviews on sampling-based posterior inference see Liu (2001); Tierney (1994). For a more theoretical treatise of transition kernels, see Feller (1971) and the monograph by Nummelin (1984). The blocked Gibbs sampler for Dirichlet process mixtures was proposed by MacEachern (1994). Dirichlet process mixture inference has been studied intensively in the literature, both in statistics and in machine learning. Neal (2000) gives a concise overview of available sampling algorithms. A more recent development in DP Markov chain sampling are split-merge samplers (Green and Richardson, 2001; Jain and Neal, 2004, 2007). Split-merge samplers result in an algorithmic behavior roughly similar to Green's reversible jump (Green, 1995). Application of reversible jump methods to mixture model order selection (without consideration of Dirichlet processes) was studied by Richardson and Green (1997). The interest of the machine learning community resulted in approximate variational inference strategies (Blei and Jordan, 2004), as well as incorporation of variational techniques into sampling processes (Teh *et al.*, 2008b).

Blocking strategies are studied explicitly in mathematical optimization, but have also been identified rather independently in the MCMC literature. Many advanced sampling algorithms are essentially blocking schemes. These include the multigrid Monte Carlo simulators of Goodman and Sokal (1989), where blocks are defined by grouping observations on a grid, and to some degree the coarse-scale chains of Higdon *et al.* (2003), as well as the group move strategies in Liu and Sabatti (1998) (though in the latter case, the term “group” refers to an algebraic group rather than a collection of coordinates). The survey of Hunter and Lange (2004) is a frequently cited source on majorization in EM-style algorithms, but the far more general perspective taken by de Leeuw (1994) and Heiser (1995) is perhaps more useful.

2.7 Markov Random Fields

Markov random fields provide an approach to the difficult problem of modeling systems of dependent random variables. To reduce the complexity of the problem, interactions are restricted to occur only within small groups of variables. Dependence structure can conveniently be represented by a graph, with vertices representing random variables and an edge between two vertices indicating statistical dependence. More formally, a MRF is a collection of random variables defined on an undirected, weighted graph $\mathcal{N} = (V_{\mathcal{N}}, E_{\mathcal{N}}, W_{\mathcal{N}})$, the *neighborhood graph*. The vertices in the vertex set $V_{\mathcal{N}} = \{v_1, \dots, v_n\}$ are referred to as *sites*. $E_{\mathcal{N}}$ is the set of graph edges, and $W_{\mathcal{N}}$ denotes a set of constant edge weights. Since the graph is undirected, the edge weights $w_{ij} \in W_{\mathcal{N}}$ are symmetric ($w_{ij} = w_{ji}$). Each site v_i is associated with an observation $\mathbf{x}_i$ and a random variable θ_i. When dealing with subsets of parameters, we will use the notation $\theta_A := \{\theta_i | i \in A\}$ for all parameters with indices in the set A. In particular, $\partial\,(i) := \{j | (i, j) \in E_{\mathcal{N}}\}$ denotes the index set of neighbors of v_i in $\mathcal{N}$, and $\theta_{-i} := \{\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n\}$ is a shorthand notation for the parameter set with θ_i removed.

Markov random fields model constraints and dependencies in Bayesian spatial statistics. A joint distribution Π on the parameters $\theta_1, \dots, \theta_n$ is called a Markov random field w. r. t. $\mathcal{N}$ if

$$\Pi(\theta_i|\theta_{-i}) = \Pi(\theta_i|\theta_{\partial(i)}) \qquad \text{for all } v_i \in V_{\mathcal{N}}\ . \tag{2.7.1}$$

This *Markov property* states that the random variables θ_i are dependent, but dependencies are local, i. e. restricted to variables adjacent in the graph $\mathcal{N}$. The MRF distribution $\Pi(\theta_1, \dots, \theta_n)$ plays the role of a prior in a Bayesian model. The random variable θ_i describes a parameter for the generation of the observation $\mathbf{x}_i$. Parameter and observation at each site are linked by a parametric likelihood F, i. e. each $\mathbf{x}_i$ is assumed to be drawn $\mathbf{x}_i \sim F(\,.\,|\theta_i)$. For the image processing application discussed in Sec. 4.1.4, each site corresponds to a location in the image; two sites are connected by an edge in $\mathcal{N}$ if their locations in the image are adjacent. The observations $\mathbf{x}_i$ are local image features extracted at each site.

Defining a MRF distribution to model a given problem requires verification of the Markov property (2.7.1) for all conditionals of the distribution, a tedious task even for a small number of random variables and often infeasible for large systems. The Hammersley-Clifford theorem (Besag, 1974) provides an equivalent property which is easier to verify. The property is formulated as a condition on the MRF *cost function*, and is particularly well-suited for modeling. A cost function is a function $H : \Omega_\theta^n \to \mathbb{R}$ of the form

$$H\left(\theta_1, \ldots, \theta_n\right) := \sum_{A \subset V_{\mathcal{N}}} H_A(\theta_A) \; . \qquad (2.7.2)$$

The sum ranges over all possible subsets A of nodes in the graph $\mathcal{N}$. On each of these sets, costs are defined by a local cost function H_A, and θ_A denotes the parameter subset $\{\theta_i | v_i \in A\}$. The cost function H defines a distribution by means of

$$\Pi(\theta_1, \ldots, \theta_n) := \frac{1}{Z_H} \exp(-H(\theta_1, \ldots, \theta_n)) \; , \qquad (2.7.3)$$

with a normalization term Z_H (the *partition function*). The cost function H can once again be interpreted as an energy, and (2.7.3) is an energy representation similar to (2.2.9). Without further requirements, this distribution does not in general satisfy (2.7.1). By the Hammersley-Clifford theorem, the cost function (2.7.2) will define a MRF if and only if

$$H\left(\theta_1, \ldots, \theta_n\right) = \sum_{C \subset \mathcal{C}} H_C(\theta_C) \; , \qquad (2.7.4)$$

where $\mathcal{C}$ denotes the set of all cliques, or completely connected subsets, of $V_{\mathcal{N}}$. In other words, the distribution defined by H will be a MRF if the local cost contributions H_A vanish for every subset A of nodes which are not completely connected. In other words, coupling in MRFs is direct. If two nodes are not connected by a direct edge, their joint behavior does not affect the distribution. Defining MRF distributions therefore comes down to defining a proper cost function of the form (2.7.4).

Inference algorithms for MRF distributions rely on the full conditional distributions

$$\Pi(\theta_i | \theta_{-i}) = \frac{\Pi(\theta_1, \ldots, \theta_n)}{\int \Pi(\theta_1, \ldots, \theta_n) d\theta_i} \; . \qquad (2.7.5)$$

For sampling or optimization algorithms, it is typically sufficient to evaluate distributions up to a constant coefficient. Since the integral in the denominator is constant with respect to θ_i, it may be neglected, and the full conditional can be evaluated for algorithmic purposes by substituting given values for all parameters in θ_{-i} into the functional form of the joint distribution $\Pi(\theta_1, \dots, \theta_n)$. Due to the Markov property (2.7.1), the full conditional for θ_i is completely defined by those components H_C of the cost function for which $i \in C$. This restricted cost function will be denoted $H(\theta_i|\theta_{-i})$. The set $C \subset V_{\mathcal{N}}$ of nodes over which $H(\theta_i|\theta_{-i})$ is effectively computed is also known as the *Markov blanket* of i. A coordinate or block optimization method with target function H can solve the individual optimization problems (2.6.2) for each node i by optimizing a term computed over the Markov blanket of i. If dependencies are sufficiently local, these sets are small, and optimization is feasible. The most popular choice of coordinate optimization schemes in the MRF literature are Gibbs samplers.

A simple example of a MRF cost function with continuously-valued parameters θ_i is

$$H\left(\theta_i|\theta_{-i}\right) := \sum_{l \in \partial(i)} \|\theta_i - \theta_l\|^2 . \tag{2.7.6}$$

The resulting conditional prior contribution $M(\theta_i|\theta_{-i}) \propto \prod_{l \in \partial(i)} \exp(-\|\theta_i - \theta_l\|^2)$ will favor similar parameter values at sites which are neighbors.

In the case of clustering problems, the constraints are modeled on the discrete set $\{Z_1, \dots, Z_n\}$ of class label indicators. Cost functions such as (2.7.6) are inadequate for this type of problem, because they depend on the magnitude of a distance between parameter values. If the numerical difference between two parameters is small, the resulting costs are small as well. Cluster labels are not usually associated with such a notion of proximity: Most clustering problems (with some exceptions, such as Kohonen maps) do not define an order on class labels, and two class labels are either identical or different. This binary concept of similarity is expressed by cost functions such as

$$H(\theta_i|\theta_{-i}) = -\lambda \sum_{l \in \partial(i)} w_{il} \delta_{Z_i, Z_l} , \tag{2.7.7}$$

where δ is the Kronecker symbol, λ a positive constant and w_{il} are edge weights. The class indicators Z_i, Z_l specify the classes defined by the parameters θ_i and θ_l. Hence, if θ_i defines a class different from the classes of all neighbors, $\exp(-H) = 1$, whereas $\exp(-H)$ will increase if at least one neighbor is assigned to the same class. More generally, we consider cost functions satisfying

$$\begin{aligned} H(\theta_i|\theta_{-i}) &= 0 \quad \text{if } Z_i \notin Z_{\partial(i)} \\ H(\theta_i|\theta_{-i}) &< 0 \quad \text{if } Z_i \in Z_{\partial(i)} . \end{aligned} \tag{2.7.8}$$

The function will usually be defined to assume a larger negative value the more neighbors are assigned to the class defined by θ_i. Such a cost function may be used, for example, to express smoothness constraints on the cluster labels, as they encourage smooth assignments of adjacent sites. In Bayesian image processing, label constraints may be used to smooth the results of segmentation algorithms, as first proposed by Geman et al. Geman *et al.* (1990).

2.7.1 References

Graph-based interaction models and “lattice systems” have a long history in physics, in the form of the Ising model, Potts model and related structures. Long before these models were recognized in statistics, physicists had already developed a number of inference techniques of variational type, notably the mean-field method. Cipra (1987) gives an overview. Statistical physics studies these models in terms of Gibbs distributions, and the link between statistical physics and spatial statistics was established by a proof of equivalence of Gibbs distributions and Markov random fields. Hammersley and Clifford (1971) found such a proof of what is now known as the Hammersley-Clifford theorem, but did not publish it at the time. Nonetheless, the result quickly became known in the field, and when Besag (1974) simplified and published the proof three years later in a discussion paper, both the paper and the discussion contributions already take the name “Hammersley-Clifford theorem” for granted. Besag and Green (1993) credit Grenander (1983) with pioneering the Bayesian approach to spatial statistics as used today throughout statistics and computer vision. Markov random field methods became practically feasible with the introduction of the Gibbs sampling algorithm for random fields by Geman and Geman (1984). Gaussian random fields, for which the adjacency matrix of the random field graph corresponds to the inverse covariance matrix of a joint Gaussian distribution on the graph nodes, were suggested by Speed and Kiiveri (1986). The idea of graph edge weights determined by the observed data was introduced in Geman *et al.* (1990). Literature surveys include Besag *et al.* (1995), and the textbook of Winkler (2003). Combination of Markov-type interaction models and multiresolution approaches are covered by Willsky (2002).

Chapter 3

Clustering: Parametric Mixtures

Before considering clustering models of nonparametric Bayesian type in Chap. 4, we will discuss two finite mixture models in this chapter, which somewhat differ from well-known standard models in terms of their components. One clusters ranking data, i. e. preference lists represented as permutations, by means of a parametric family of distributions on the symmetric group. The second model is a mixture-of-mixtures model, to address the problem of clustering speckle noise data which arises in SAR image segmentation problems.

Acknowledgment. *All research presented in this chapter was conducted under supervision of Joachim M. Buhmann. The material presented in Sec. 3.1 is the result of research conducted in cooperation with Ludwig M. Busse, and the heterogeneity problem emerged in the course of our discussions. His further contributions include the implementation of the model; its experimental evaluation; and an extensive literature review prior to the actual research, on which our subsequent work was based. My own contributions include the model and its derivation; the derivation of the algorithm (though the local search in the M-step, which decisively increases algorithmic efficiency, was suggested by Ludwig); and the text of the original publication, which is reproduced here with minor modifications.*

3.1 Cluster Analysis of Heterogeneous Rank Data

The term *rank data* refers to data in which each measurement is a ranking, an arrangement of a given set of items in order of preference. Rankings occur in consumer questionnaires, voting forms or other inquiries of preferences. Cluster analysis of rank data attempts to identify typical groups of rank choices. Available clustering methods are based on mixtures of parametric distributions on the symmetric group. Empirically measured rankings are often incomplete, i. e. different numbers of filled rank positions cause heterogeneity in the data. This section proposes a mixture model for clustering of heterogeneous rank data. Rankings of different lengths can be described and compared by means of a single probabilistic model. A maximum entropy approach avoids hidden assumptions about missing rank positions. Parameter estimators and an efficient EM algorithm for unsupervised inference are derived for the ranking mixture model. Experiments on both synthetic data and real-world data demonstrate significantly improved parameter estimates on heterogeneous data when the incomplete rankings are included in the inference process.

3.1.1 Introduction

Ranking data commonly occurs in preference surveys: A number of subjects are asked to rank a list of items or concepts according to their personal order of preference. Two types of ranking data are usually discussed in the literature: Complete and partial (or incomplete) rankings. A wide range of probabilistic models is available for both (Diaconis, 1988; Critchlow, 1985). A complete ranking of r items is a permutation of these items, listed in order of preference. Mathematical models of rankings are based on the corresponding permutation group. A partial ranking is a preference list of t out of r items. Partial rankings require some refinements of models designed for complete rankings, since two arbitrary partial rankings will in general contain different subsets of the items. An extensive review of rank comparisons can be found in (Critchlow, 1985).

Clustering of rank data aims at the identification of groups of rankers with a common, typical preference behavior (Marden, 1995). An unsupervised clustering method for complete rankings has been proposed by Murphy and Martin (2003), based on the well-known Mallows' model (Mallows, 1957) and its generalizations. A different but related problem is the combination of several rankings. This question has recently been discussed by a number of authors, both in machine learning (Lebanon and Lafferty, 2002) and discrete algorithmics (Ailon *et al.*, 2005).

For real-world surveys, the data analyst is often confronted with *heterogeneous* data, that is, data containing partial rankings of different lengths. In the well-studied APA data set (Diaconis, 1989), for example, only about a third of the rankings are complete, and the remaining incomplete lists have variable lengths. Common practice in the analysis of heterogeneous rank data is to delete partial rankings, and analyze only the subset of complete rankings (Murphy and Martin, 2003), or to analyze partial rankings of different lengths separately. This raises conceptual problems, as we must expect the removal of a subsample of common characteristic (i. e. incompleteness of the rankings) to cause a systematic bias. Moreover, decreasing the sample size by removing partial rankings can result in a significant decrease of estimation accuracy.

For heterogeneous data, clusters model typical preferences. A ranker associated with any group may either state his preferences completely or incompletely. In other words, each cluster again constitutes a heterogeneous data set, containing rankings of different lengths. The model introduced below builds on the work of Fligner and Verducci (1986) and is applicable to heterogeneous data. It is a parametric location-scale model based on the Kendall distance (Kendall, 1938), and thus related to the model of Mallows (1957). We address the clustering problem by combining several model instances into a parametric mixture. Inference is conducted in a maximum likelihood framework by an expectation-maximization algorithm. The model admits an estimation procedure much more efficient than the straightforward EM approach proposed in the literature for distance-based rank models. Our experiments clearly demonstrate that the additional information in partial rankings can significantly improve parameter estimates of mixture components in rank cluster analysis.

3.1.2 Background

The objective of rank data clustering is to (i) group similar rankings in the input data and (ii) identify rankings that are prototypical representatives for each group. Our approach is probabilistic: A probability model is defined capable of representing an individual group. A mixture of such models is then fitted to the data by an alternating estimation procedure. We will first introduce the standard probability models on rank data available in the literature.

Models for Complete Rankings

We assume that rank data for r items are observed. The items are indexed $m = 1, \dots, r$, and n subjects are asked to arrange the items according to their order of preference. Each of the resulting lists can be regarded as a permutation π_i of the item indices, i. e. $\pi_i(m) = j$ indicates that the i-th ranker has assigned rank j to item m. The set of possible rankings is then given by the set of possible permutations of r items. This set has a group structure and is referred to as the *symmetric group* of order r, denoted $\mathbb{S}(r)$.

Statistics has developed a sizable amount of rank data models. Of particular interest for data clustering are the so-called *distance-based models* of the form

$$P(\pi|\lambda, \sigma) := \frac{1}{Z(\lambda)} \exp\left(-\lambda d(\pi, \sigma)\right) \ , \tag{3.1.1}$$

with $Z(\lambda) := \sum_{\pi \in \mathbb{S}(r)} \exp\left(-\lambda d(\pi, \sigma)\right)$. The model is parameterized by a ranking $\sigma \in \mathbb{S}(r)$ and a dispersion parameter $\lambda \in \mathbb{R}_+$. The function $d : \mathbb{S}(r) \times \mathbb{S}(r) \to \mathbb{R}_{\geq 0}$ is a *distance function*, i. e. a function with metric properties on $\mathbb{S}(r)$. Since d is a metric and hence $d(\pi, \sigma) = 0$ iff $\pi = \sigma$, the distribution P assumes its unique mode at σ, and σ is referred to as the *modal ranking*. The dispersion parameter λ controls how sharply the distribution peaks around the mode, i.e. small (large) λ values code for broad (peaked) distributions. For clustering, distance-based models capture the notion that two observations belong to the same group if they are "close". The approach is related to familiar clustering methods for other data types, such as Gaussian mixtures for vectorial data (which measure distance by Euclidean or covariance-adjusted Euclidean distance) or multinomial mixtures for histogram data (which measure a distance-like quantity by Kullback-Leibler divergence). Different models can be obtained by substituting different types of metrics for d in (3.1.1). Other popular choices include the Spearman rank correlation metric, and the Hamming, Cayley and Ulam distances (Critchlow, 1985). The present work focuses on one metric in particular, the widely used *Kendall distance* (Kendall, 1938), defined as

$$\begin{aligned} d_\tau(\pi, \sigma) := \ &\text{minimum of adjacent transpositions required to} \\ &\text{transform } \pi \text{ into } \sigma. \end{aligned} \tag{3.1.2}$$

Closely related is the *Cayley distance*, which drops the adjacency requirement, and thus measures the distance in terms of arbitrary transpositions. For $d = d_\tau$, the model (3.1.1) is *Mallows' ϕ model* (Mallows, 1957) in its original form. More generally, models of the form (3.1.1) are usually referred to as Mallows models, provided that d is a metric.

Clustering with Mallows' Model

For clustering, the observed rank data is assumed to consist of K groups. Each group is modeled by a Mallows distribution

$$P_k(\pi|\lambda_k, \sigma_k) := \frac{1}{Z(\lambda_k)} \exp\left(-\lambda_k d_\tau(\pi, \sigma_k)\right) \ . \tag{3.1.3}$$

The component distributions are joined in a mixture model,

$$Q(\pi) := \sum_{k=1}^{K} c_k P_k(\pi|\lambda_k, \sigma_k) \ , \tag{3.1.4}$$

where the mixture weights $(c_1, \ldots, c_K)$ form a partition of 1. Model parameters can be estimated with an expectation-maximization (EM) algorithm (McLachlan and Krishnan, 1997), or more sophisticated latent variable estimation algorithms.

Partial Rankings

A *partial ranking* is a ranking of t out of r items. Usually, one assumes a top-t ranking, i. e. subjects have ranked their t favorites out of a larger number of r items. Distance-based models for partial rankings can be constructed by generalizing metrics on complete rankings to valid metrics on partial rankings. (Critchlow, 1985) has proposed such a generalization based on Hausdorff distances.

A partial top-t ranking is best represented as an inverse: In standard notation, regarding the permutation π as a list of numbers, position in the list corresponds to an item index (and the entry value at that position gives a rank). A ranking of t favorite items is thus a list with gaps. Written as the inverse π^{-1}, position denotes rank, and a top-t ranking has the form $\pi^{-1} = (\pi^{-1}(1), \ldots, \pi^{-1}(t), *, \ldots, *)$. For any partial ranking π of length t, denote by $C(\pi)$ the set of all complete rankings $\tilde{\pi}$ matching π in their first t positions, that is, $C(\pi) := \{\tilde{\pi} \in \mathbb{S}(r) | \tilde{\pi}(j) = \pi(j), j = 1, \ldots, t\}$. We will refer to C as the *consistent set* of π (in algebraic terms, this is just the right coset $\mathbb{S}_{r-t}\pi$). For any two different partial rankings of the same length, the consistent sets are disjoint, and their union over all partial rankings of a given length is $\mathbb{S}(r)$. For a given metric d on $\mathbb{S}(r)$, Critchlow (1985) defines an induced metric d^* on partial rankings as the Hausdorff distance between their consistent sets. As put by Critchlow, $d^*(\pi, \sigma)$ can be imagined as the smallest amount by which $C(\pi)$ has to be enlarged to include all of $C(\sigma)$. Another approach to partial rankings is the completion method proposed by Beckett (1993), who estimates complete rankings from partial ones based on a Mallows model (cf. Sec. 3.1.5).

3.1.3 Modeling Heterogeneous Data

In the present work, we consider the problem of modeling real-world survey data, which usually includes partial rankings of variable length t. Differences arise because many subjects will rank only their favorite t items. For ranking data on r items, we therefore have to assume an observed sample to contain partial rankings of all possible lengths $t = 1, \ldots, (r-1)$ (note that $t = (r-1)$ is equivalent to $t = r$, since the missing position is uniquely determined). [1]

Choice of Metric

The model described in this section is based on the Kendall distance. Our choice of the metric is motivated by a range of properties: First, it has an intuitive and plausible interpretation as a number of pairwise choices. Mallows (1957) argues that it provides the best possible description of the process of ranking items as performed by a human. Second, it enjoys a high de-facto relevance due to its widespread use. Third, there are a number of appealing mathematical properties: It counts (rather than measures), is efficiently computable, decomposable into a sum, and its standardized distribution has a normal limit (Diaconis, 1988). Though our study is limited to the Kendall case, Fligner-Verducci type models can be derived for the Cayley distance as well (Fligner and Verducci, 1986).

Probabilistic Model

If only a subset of the available items is ranked, the choice of a probabilistic model implies a distribution assumption for the missing entries. We take a maximum entropy approach, demanding our model to be maximally noncommittal with respect to the missing information. Such a model is suitable to address several generative scenarios for partial rankings: One is indifference of the ranker, i. e. a subject ranks t favorite items, but does not have any preferences concerning the remainder. Another setting are large sets of items, where most subjects will not take the time to provide a complete list (e. g. when the task is to specify a ranking of favorites out of thousands of items). In general, the approach is applicable unless prior information on the popularity of items is available. A maximum entropy approach attempts to avoid implicit (hidden) assumptions on the choice of items. This is a notable difference to the Hausdorff metric approach, for

[1] We do not consider partial rankings with gaps, i. e. rankings with a total of $t < r$ filled position and empty ranks in between, since data of this type can be expected to be rare. Our model does, in principle, generalize to the case of rankings with gaps, but the actual computations become more difficult.

example, which constitutes a worst-case assumption: The distance problem is reduced to the original metric by expanding a pair of partial rankings into that consistent pair of complete rankings which differs most under the inducing metric.

To express lack of knowledge w. r. t. to items beyond the preferred t choices, we have to assume that the ranker's choice effectively encompasses all possible completions of π to a complete ranking in $\mathbb{S}(r)$. In other words, successive ranking of items is regarded as a constraining process: By each additional item entered into the list, the ranker constrains the set of possible completions. A full ranking limits $\mathbb{S}(r)$ down to a single element. A partial ranking defines the set $C(\pi)$ of possible completions. Any model distribution F on complete rankings can then be generalized to a distribution F^t on partial rankings by defining the probability of π under F^t as the total probability placed on the set $C(\pi)$ by the model F:

$$F^t(\pi) := F(C(\pi)) = \sum_{\tilde{\pi} \in C(\pi)} F(\tilde{\pi}) \ . \tag{3.1.5}$$

For Mallows' model based on the Kendall distance, the probability $F(C(\pi))$ admits an elegant decomposition. From a statistics point of view, the approach can be regarded as a censored data problem. For the Kendall metric, censored rank data has been considered in (Fligner and Verducci, 1986). They build on the well-known fact that the Kendall distance, as well as the Cayley and Hamming distances, can be decomposed into a sum. Define the following statistic for each position $j = 1, \ldots, (r-1)$ in a complete ranking π of length r:

$$\tilde{s}_j(\pi) := \sum_{l=j+1}^{r} \mathbb{I}\{\pi^{-1}(j) > \pi^{-1}(l)\} \ , \tag{3.1.6}$$

where π^{-1} denotes the inverse of π in $\mathbb{S}(r)$ and I the indicator function of a set. Intuitively, $\tilde{s}_j$ is the number of adjacent transpositions required to move item j to position j, if the items at the previous $1, \ldots, (j-1)$ are already ordered. The sum over the statistics $\tilde{s}_j$ is the Kendall distance of π and the identity permutation $\mathrm{Id}_{\mathbb{S}(r)}$ (Fligner and Verducci, 1986). The metric d_τ is *right-invariant*, that is, for any $\pi_1, \pi_2, \pi_3 \in \mathbb{S}(r)$, $d_\tau(\pi_1\pi_3, \pi_2\pi_3) = d_\tau(\pi_1, \pi_2)$. Hence, for any $\sigma \in \mathbb{S}(r)$,

$$d_\tau(\pi, \sigma) = d_\tau(\pi\sigma^{-1}, \mathrm{Id}_{\mathbb{S}(r)}) = \sum_{j=1}^{r-1} \tilde{s}_j(\pi\sigma^{-1}) \ . \tag{3.1.7}$$

This representation is somewhat inconvenient for modeling partial rankings, since the sum ranges over the suffix of rank j, which includes empty positions. We therefore substitute equivalent statistics s_j involving only indices up to j. For any permutation ρ, define

$$s_j(\rho) := \rho(j) - \sum_{l=1}^{j} \mathbb{I}\{\rho(j) \geq \rho(l)\} \; . \tag{3.1.8}$$

The Kendall metric is then computed as $d_\tau(\pi, \sigma) := \sum_{j=1}^{r} s_j(\sigma\pi^{-1})$, which avoids any explicit use of π: Since π^{-1} is a top-t list, it is not invertible. The importance of the sum representation for modeling partial rankings is that it can be decomposed into terms corresponding to filled and empty positions, respectively:

$$d_\tau(\pi, \sigma) = \sum_{j=1}^{t} s_j(\sigma\pi^{-1}) + \sum_{j=t+1}^{r} s_j(\sigma\pi^{-1}) = \mathbf{s}^t(\sigma\pi^{-1}) + \mathbf{s}^{\text{empty}}(\sigma\pi^{-1}) \tag{3.1.9}$$

The probability of the consistent set of π under Mallows' model can then be expressed as

$$\begin{aligned} F(C(\pi)|\lambda, \sigma) &= \frac{1}{Z(\lambda)} \sum_{\tilde{\pi} \in C(\pi)} \exp\left(-\lambda d_\tau(\tilde{\pi}, \sigma)\right) \\ &= \frac{\exp\left(-\lambda \mathbf{s}^t(\sigma\pi^{-1})\right)}{Z(\lambda)} \sum_{\tilde{\pi} \in C(\pi)} \exp\left(-\lambda \mathbf{s}^{\text{empty}}(\sigma\tilde{\pi}^{-1})\right) \end{aligned} \tag{3.1.10}$$

The sum over $C(\pi)$ depends only on t, and is absorbed into the partition function $Z(\lambda)$. Hence, the resulting partition function $Z^t(\lambda)$ depends on t. The probability of the partial ranking is thus

$$F(C(\pi)|\lambda, \sigma) = \frac{1}{Z^t(\lambda)} \exp\left(-\lambda \mathbf{s}^t(\sigma\pi^{-1})\right) \; , \tag{3.1.11}$$

and we write $F(\pi|\lambda, \sigma) := F(C(\pi)|\lambda, \sigma)$. The partition function Z^t can be derived from the (somewhat more complicated) model in (Fligner and Verducci, 1986), as

$$Z^t(\lambda) := \prod_{j=1}^{t} \frac{1 - e^{-\lambda(r-j+1)}}{1 - e^{-\lambda}} \; . \tag{3.1.12}$$

The distribution is a maximum entropy model, as it constitutes an exponential family distribution given the modal ranking σ, with the functions s_j as its sufficient statistics. The choice of the location parameter σ does not change the model's entropy.

Heterogeneous, partial ranking data drawn from K distinct groups can now be described by a mixture model. Denote by $t(\pi)$ the length of an arbitrary partial ranking π. The generative model for the data is then

$$Q(\pi|\mathbf{c}, \boldsymbol{\lambda}, \boldsymbol{\sigma}) := \sum_{k=1}^{K} \frac{c_k}{Z^{t(\pi)}(\lambda_k)} e^{-\lambda_k \mathbf{s}^{t(\pi)}(\sigma\pi_k^{-1})} \ . \tag{3.1.13}$$

To summarize, lack of knowledge (or indifference of a ranker) about the order of neglected items is expressed by substituting the consistent set of a ranking in the modeling process. Probabilities are comparable for rankings of different lengths. Formally, this holds because the model is a distribution on the consistent sets $C(\pi)$. For any two rankings, the sets are nested if one ranking prefixes the other, and are disjoint otherwise. The mixture expresses the separation of the rankers surveyed in the data into different groups or types, each of which exhibits a "typical" preference behavior. The data collected from rankers within a single group will in general be heterogeneous. For a given group, the modal ranking describes a consensus preference, and the corresponding dispersion parameter variation between the associated rankers.

3.1.4 Model Inference

Our approach to inference is based on maximum likelihood (ML) estimation. For the mixture model described above, the overall ML estimator of the model parameters is approximated with an expectation-maximization (EM) algorithm (McLachlan and Krishnan, 1997). In this section, we derive estimation equations for the heterogeneous data model, and discuss the implementation of efficient EM algorithms for rank data. Straightforward implementations of such algorithms previously proposed for Mallows mixtures on complete rankings (Murphy and Martin, 2003) require the repeated evaluation of sums over all possible rankings. Since the symmetric group $\mathbb{S}(r)$ has $r!$ elements, such methods are only applicable for rankings with a small number of entries.

For data $\pi_i, i = 1, \ldots, n$ and K clusters, we define binary class assignment vectors $\mathbf{Z}_i := (Z_{i1}, \ldots, Z_{iK})$. If π_i is assigned to cluster k, then $Z_{ik} = 1$ and all other entries are set to zero. These are the hidden variables of the EM estimation problem. The EM algorithm relaxes the binary assignments to assignment probabilities $q_{ik} := \mathbb{E}[Z_{ik}]$, where $q_{ik} \in [0, 1]$ and $\sum_k q_{ik} = 1$ for each i. The E-step of the algorithm computes estimates of the assignment probabilities conditional on the current parameter configuration of the model. Given estimates of the component parameters

λ_k, σ_k and the mixture weight c_k for each cluster k, assignment probabilities are estimated as $q_{ik} := \frac{c_k F^t(\pi_i|\lambda_k,\sigma_k)}{\sum_{l=1}^K c_l F^t(\pi_i|\lambda_l,\sigma_l)}$. In the M-step, assignment probabilities are assumed to be given. For each cluster, the parameters to be estimated are c_k, λ_k and σ_k. As for any mixture model EM algorithm, the mixture weights are straightforwardly computed as $c_k := \frac{1}{n}\sum_{i=1}^n q_{ik}$. ML estimation of the component parameters σ_k, λ_k proceeds in two steps, first obtaining an estimate of σ_k (which does not depend on λ_k), and then estimating λ_k conditional on σ_k. This is reminiscent of e. g. the two-stage ML estimation of location and scale parameters for Gaussian models. The modal ranking ML estimate is

$$\hat{\sigma}_k = \arg\max_{\sigma_k} \log \prod_{i=1}^n F(\pi_i^t|\lambda_k,\sigma_k)^{q_{ik}} = \arg\min_{\sigma_k} \sum_{i=1}^n q_{ik} \sum_{j=1}^{t(\pi_i)} s_j(\sigma_k \pi_i^{-1}) \ . \tag{3.1.14}$$

Rather than evaluating the minimum over the whole group, our algorithm performs a local search step, by minimizing over all adjacent transpositions around the estimate $\tilde{\sigma}_k$ obtained during the previous M-step. This strategy is equivalent to searching within a d_τ-radius of 1. When initialized at random, the algorithm may thus require several steps until it reaches the correct σ_k. The local search results in a generalized EM (GEM) algorithm, since the conditional likelihood is increased but not fully maximized during the M-step. Generalized EM algorithms satisfy the EM convergence conditions and retain EM convergence guarantees (McLachlan and Krishnan, 1997). Our control experiments in Sec. 3.1.5 clearly indicate that the local estimation approach is adequate. If modal ranking estimation errors occur, they are due to ambiguous data, i. e. data drawn from clusters for which the distance between the modal rankings is small w. r. t. to their dispersion. Local search over transpositions reduces the estimation costs for σ_k from $r!$ to r evaluations.

Since the dispersion parameter is continuous, a maximum condition for the likelihood w. r. t. λ_k can be obtained by differentiation. Setting the derivative of the log-likelihood of one mode to zero yields

$$-\sum_{i=1}^n \frac{\partial}{\partial \lambda_k} \log Z^{t(\pi_i)}(\lambda_k) = \sum_{i=1}^n d(\pi_i, \sigma_k) \ . \tag{3.1.15}$$

For our heterogeneous data model as described in Sec. 3.1.3, (i) the partition function has a closed-form solution and the derivative can be obtained explicitly, and (ii) the model has to be decomposed over different types of rankings, since the partition function depends on t. Assume that the observations π_i have different lengths $t \in \{1, \ldots, r\}$. Denote by $I_t \subset \{1, \ldots, n\}$

the set of indices i for which π_i has length t. The log-likelihood of the complete data set under cluster k is

$$\begin{aligned}\log \prod_{i=1}^{n} F(\pi_i|\lambda_k, \sigma_k) &= \sum_{t=1}^{r} \sum_{i \in I_t} \log F(\pi_i|\lambda_k, \sigma_k) \\ &= -\sum_{t=1}^{r} |I_t| \log(Z^t(\lambda_k)) - \sum_{t=1}^{r} \sum_{i \in I_t} \lambda_k \sum_{j=1}^{t} s_j(\sigma_k \pi_i^{-1})\end{aligned} \tag{3.1.16}$$

Equating the derivative to zero gives

$$-\sum_{t=1}^{r} |I_t| \frac{\partial}{\partial \lambda_k} \log(Z^t(\lambda_k)) = \sum_{i=1}^{n} \sum_{j=1}^{t(\pi_i)} s_j(\sigma_k \pi_i^{-1}) \; . \tag{3.1.17}$$

The derivative of $\log(Z^t(\lambda_k))$ for given t is

$$\frac{\partial}{\partial \lambda_k} \log(Z^t(\lambda_k)) \quad = \quad \sum_{j=r-t+1}^{r} \frac{j}{e^{j\lambda_k} - 1} - \frac{t}{e^{\lambda_k} - 1} \; .$$

This expression is both rapidly computable and smooth w. r. t. λ_k. The right hand side of (3.1.17) does not depend on λ_k, hence the maximum likelihood estimator $\hat{\lambda}_k$ can be efficiently evaluated by numerical solution of equation (3.1.17).

3.1.5 Experimental Results

The experiments include artificial and real-world rank data. The mixture analysis with artificial data drawn from a density with known parameters is conducted to evaluate the algorithm's effectiveness in recovering parameters from rank data. Additional experiments are conducted on the American Psychological Association (APA) data set (Diaconis, 1989). All experiments are performed with the EM algorithm described in Sec. 3.1.4. The number of clusters is selected by a Bayesian Information Criterion (BIC)[2]

[2] In Chap. 4, model order selection is performed with a Dirichlet process. BIC is used here for two reasons: First, model order selection is not the principal focus of this work, and BIC, as the most commonly used strategy, facilitates comparison to other studies in the rank data literature. Second, the Dirichlet process is applied straightforwardly only if the model in question admits a conjugate prior. The Fligner-Verducci model is an exponential family distribution, and hence does admit a conjugate prior, but only w. r. t. the dispersion parameter.

Settings		Results					
		c uniform			c non-uniform		
d	λ	$\hat{K}$	error $\hat{c}$	error $\hat{\lambda}$	$\hat{K}$	error $\hat{c}$	error $\hat{\lambda}$
[2, 9, 9]	0.50	1	0.033	0.086	1	0.248	0.324
	1.00	3	0.007	0.056	3	0.013	0.032
	1.50	3	0.027	0.151	3	0.001	0.048
[8, 6, 6]	0.50	1	0.155	0.274	1	0.189	0.331
	1.00	3	0.029	0.094	3	0.047	0.144
	1.50	3	0.016	0.050	3	0.013	0.057

Table 3.1: Estimation errors on artificial data of sample size $n = 300$, with $K = 3$ clusters. For uniform c, all clusters have equal size. For non-uniform c, cluster sizes differ.

(McLachlan and Krishnan, 1997). For comparison, we use a clustering approach based on the completion method described in (Beckett, 1993). The method explicitly estimates a maximum likelihood completion to a full ranking by treating the missing positions as latent information, and assuming complete rankings to be distributed according to a Mallows model. An estimate of the full ranking is obtained with an EM algorithm, which alternatingly estimates a Mallows model from current completion estimates, and then estimates completions based on the current model. The method can be used as basis for partial rank data clustering model, by performing completions based on the data currently assigned to a cluster during the clustering E-step, and performing maximum likelihood estimation for the mixture components given the current completion estimates during the M-step.

Table 3.2: Long rankings: Estimation error comparison for ranking length $r = 20$, with $K = 10$ clusters and $n = 1000$ samples (uniform over partial lengths).

Method	error $\hat{\sigma}_k$	error $\hat{\lambda}_k$
Maximum Entropy	0	0.06 ± 0.01
Beckett's completion	1.52 ± 0.57	0.11 ± 0.02

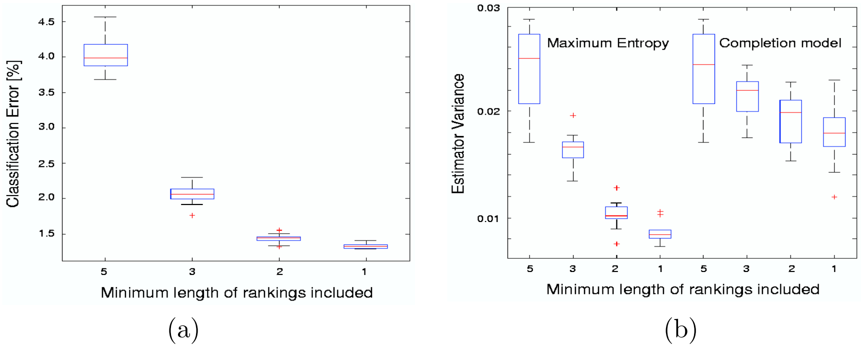


Figure 3.1: (a) Full versus restricted data set: Average estimation error for cluster assignments (vertical) versus the number of ranking types present in the data set (horizontal). (b) APA data set: Variance of dispersion estimates (vertical) versus number of ranking types present in the data set (horizontal), for our method (left) and Beckett's completion model (right). Minimum length 5 corresponds to the subset of complete rankings, 1 to the whole data set. The variance is computed over 20 bootstrap samples.

Synthetic Data

Synthetic data observations were drawn at random from a mixture model (3.1.13). Sample experiments for $r = 5$ items and $K = 3$ clusters are shown in Tab. 3.1. By d, we denote the pairwise Kendall distances between the cluster centers. The quality of parameter estimates is reported as mean squared error on $n = 300$ observations. The BIC estimate $\hat{K}$ of the number of clusters is accurate except for very small λ which corresponds to broad modes. This behavior is expected since the different modes strongly overlap for small λ and, consequently, are not resolvable for the chosen number of observations. When BIC underestimates the number of clusters, the estimation errors for λ and c generally increase. Estimation errors increase again for $\lambda = 1.5$ in the case of two close clusters ($d = [2, 9, 9]$), a distortion effect caused by points of the neighboring cluster. The dispersion at which the effect becomes visible depends on a trade-off between the dispersion and the distance of the clusters. It will occur at a larger value of λ if the clusters are closer. Remarkably, the modal rankings σ_k are always estimated correctly, unless the estimate of the cluster number is wrong.

The value of partial rankings for estimation is illustrated by Fig. 3.1.5. EM estimation of the mixture model was conducted on a random data set, with $r = 5$ and a proportion of 25% complete rankings. The partial rankings of lengths $\{1, 2, 3\}$ are also drawn with probability 0.25 each. The estimation error for the cluster assignments was recorded and plotted against the number of ranking length types present in the data (horizontal), where 5 denotes the case where all partial rankings are removed from the data set, corresponding to the common practice of analyzing only the subset of complete rankings. When more categories are added (with 1 corresponding to the complete heterogeneous data set), we observe a significant decrease in both the estimation error and its variance. A double-logarithmic plot of these results reveals an approximate scaling behavior of $\mathcal{O}(1/\sqrt{n})$. We conclude that, at least in the controlled setting of synthetically generated data, the inference procedure is capable of using the information carried by partial rankings to its advantage.

Comparisons with Beckett's completion method were conducted for rankings of length $r = 5$ and $r = 20$ on synthetic data. Parameter estimates obtained by our method are more accurate then those obtained by the completion approach. The difference is statistically significant even for $r = 5$, and becomes more pronounced as the number of items is increased. Results for $r = 20$ are reported in Tab. 3.2. Application of Beckett's method to rankings of this length requires a modification of the original algorithm. Beckett's estimation step completely enumerates the consistent set of each partial ranking, and hence scales exponentially in the number of unranked items. It can be made applicable to large rankings by substituting a sampling step, at the price of an increase in the variance of estimates. The completion method introduces an error in the estimation of the modal ranking. Errors are caused by the large number of latent variables required by the completion model, which result in diffuse distributions of the cluster assignments.

APA Data

The APA data set of real-world rankings was obtained from the results of the American Psychological Association's 1980 presidential election. Each ballot is a ranking of five candidates. The data set is remarkably large (about 15,000 observations) and it has been extensively analyzed (Diaconis, 1988). The data is heterogeneous, that is, only 5738 ballots contain complete rankings. The remainder contains top-t rankings of all possible lengths $t = 1$ through $t = 3$ (note that $t = 4$ is equivalent to a complete ranking). Since no ground-truth is available for this data, the estimation errors cannot be computed. However, to analyze the value of the partial rankings for estimation accuracy, we consider the variance of the estimate of λ. Fig. 3.1.5 shows a plot of the bootstrap variance estimate of the estimators $\lambda_1, \ldots, \lambda_K$, for both our model and clustering based on Beckett's completion approach. The variance estimates are plotted versus the number of ranking types (i. e. different lengths). The error bars measure variances over multiple repetitions of the bootstrap estimation experiment. For our maximum entropy model (left), inclusion of additional partial observations in the analysis clearly stabilizes parameter estimates. The variance remains notably higher for the Beckett approach (right). Using Beckett's completion requires latent variables to account for the missing positions, in addition to the assignment variables required by the mixture model. Since additional latent variables increase the overall entropy of the model, the completion approach has a destabilizing effect, which becomes more pronounced as the proportion of partial rankings in the data increases. It will also slow down convergence of the inference algorithm, as the convergence speed of EM algorithms depends on the proportion of latent variables (McLachlan and Krishnan, 1997).

3.1.6 Discussion

We have presented an unsupervised clustering approach for ranking data that is capable of performing an integrated analysis on heterogeneous, real-world data, rather then decimating the data to fit the model. An efficient EM algorithm has been derived and shown to recover parameters accurately from data.

Our method offers two advantages compared to rank data clustering techniques available in the literature: (i) the ability to analyze a data set composed of different ranking types, and (ii) efficient inference. The value of the former point was demonstrated by our experiments: Removing partial rankings from a given data set significantly reduces the accuracy of parameter estimates. For data containing only complete rankings, a decrease in estimation accuracy would have to be expected if samples are removed. That the same effect is observable (Fig. 3.1.5) when the removed rankings are partial shows that incomplete rankings carry valuable information – even those containing only a single entry.

However, on real-world survey data, this effective loss in sample size is not the only consequence of removing data. In a survey, ranking only partially may constitute a typical behavior. That is, if providing a partial rather than a complete ranking correlates with certain preferences, removing partial rankings will exclude these modes of behavior from the analysis. In addition to reducing the sample size, it also introduces a systematic bias. Both drawbacks can be avoided by automatic analysis methods capable of processing heterogeneous data, and combining estimate contributions obtained on rankings of different lengths in a meaningful way. Our modeling approach permits the natural integration of different length types by defining a distribution on the subset of completions consistent with a given partial ranking.

Algorithmic inference of our model is substantially more efficient than the algorithms available in the literature for distance-based models. The EM algorithm presented in Sec. 3.1.4 scales linearly in the number of ranked items (i. e. the order r of the permutation group), rather than exponentially, as other algorithms do (Murphy and Martin, 2003).

Our modeling approach relies on the decomposition of the Kendall distance into a sum over ranking positions and, therefore, it generalizes to ranking metrics with the same property. Such a decomposition is known for the Kendall, Cayley and Hamming distances, but results from Weyl group theory suggest that it does not exist for other metrics (Diaconis, 1988). Approximate decompositions for other metrics, however, might render efficient relaxations possible which would generalize our approach to these cases. Our emphasis on the Kendall metric is motivated by its ubiquitous usage in rank mixture analysis and by its natural properties (see Sec. 3.1.3) for rank comparisons.

3.2 Mixture-of-Mixture Models for Speckle Noise Degradation

The setting of the clustering problem studied in the following is the image segmentation problem, in the particular case of SAR (synthetic aperture radar) images. The dominant noise type in such images is multiplicative speckle noise, which is due to interference caused by radar backscatter. A number of different, analytically derived parametric models is available for such data. Their common properties include restriction to the positive semi-axis, asymmetry (positive skewness), and a leptokurtic (heavier than Gaussian) upper tail. Proposed models include gamma distributions, K distributions, Rayleigh distributions, and a host of generalizations and model combinations (see e.g. Oliver and Quegan, 1998, for an overview). In practical applications, these models face a common problem: Differences in fitting behavior between different models are minute, whereas SAR data properties can vary heavily from image to image, due to properties of the source, as well as often heavy preprocessing.[3] Such preprocessing frequently (but not always) involves various transforms of logarithmic type, turning multiplicative speckle noise into additive noise, but at the same time inverting the skew of the data distribution.

[3] Visualizing raw SAR measurements as images requires computer preprocessing to begin with, such that there is no well-defined notion of an unprocessed original image.

In Chap. 4, segmentation of images (including SAR images) will be based on histogram representations of the data, modeled by means of multinomial component distributions. In the remote sensing community, smooth parametric distributions on the real line are preferred to the inherently more flexible multinomials. The model presented here attempts to strike a balance between smoothness and flexibility. It is based on the observation that, regardless of the preprocessing involved, SAR data tends to be unimodal within segments. Unimodal distributions can be approximated well by mixtures of two or three Gaussians. We therefore use such a mixture model to represent *each* image segment, resulting in Gaussian mixture-of-mixture representation for the overall image.

Mixture-of-mixtures models, i. e. mixture models for which each component again constitutes a parametric mixture, have been studied repeatedly in the statistics and machine learning literature, for example by McLachlan and Gordon (1989). Applications of related models in supervised settings have been considered by Jordan and Jacobs (1994); Hastie and Tibshirani (1996). Supervision information simplifies the inference problem, because estimation can be conducted separately for each group. Gaussian mixture-of-mixtures models have previously been applied to image segmentation in (Hermes *et al.*, 2002), where the model is optimized by deterministic annealing (Rose, 1998), and model components are coupled between clusters to decrease computational complexity. We will derive an nested algorithm of EM type, with a blocking structure adapted to the hierarchical structure of the model. The algorithm substantially simplifies inference, without resorting to mode coupling heuristics, when the number of modes per inner mixture is small.

3.2.1 Segmentation approach

Image data is assumed to be given in form of local histograms, that is, the features extracted from the image are histogram representations of the local data distributions in the neighborhood of image pixels. The histograms are grouped into a pre-specified number of clusters, each of which is modeled by a parametric mixture model.

For a grayscale input image, a local histogram is extracted from the image at the sites (nodes) of an equidistant grid. The local histogram at a given grid point is extracted by centering a window at the respective pixel, selecting all pixels within the window and sorting their grayscale values into a histogram. This procedure results in a set of histograms $\mathbf{h}_i = (n_{i1}, \ldots, n_{iN_{\text{bins}}})$. Here $i = 1, \ldots, n$ indexes the grid points and N_{bins} is the number of histogram bins, so n_{ij} denotes the counts in bin j of histogram

i. We assume that all histograms contain an identical total number N_{counts} of counts.

The data is modeled by a *mixture-of-mixtures model*, i. e. a finite mixture model the component densities of which are themselves represented by finite mixtures. All component mixture densities consist of an identical number N_{Modes} of Gaussian components:

$$p\left(x|\theta\right) = \sum_{\tau=1}^{K} c_\tau p_\tau\left(x\right) = \sum_{\tau=1}^{K} c_\tau \Big(\sum_{\alpha=1}^{N_{\text{Modes}}} c_\alpha^\tau g_\alpha^\tau\left(x\right) \Big) , \tag{3.2.1}$$

where $g_\alpha^\tau\left(x\right) = g\left(x|\mu_\alpha^\tau, \sigma_\alpha^\tau\right)$ denotes a normal density and θ the full set of Gaussian parameters. c_τ, c_α^τ are the priors of the segments and the modes, respectively. We expect the local image histograms to be uni- or at most bimodal, so we are interested only in cases where the number of inner components is small (typically $N_{\text{Modes}} = 2, 3$).

Since the range of digital image data is restricted to a finite intensity interval, we have to truncate the Gaussians. These distributions are referred to as *rectified distributions* in the literature (Socci *et al.*, 1998). Rectification somewhat complicates parameter estimation, because a ML estimator for a Gaussian mean or variance parameter is not a valid ML estimator for the rectified Gaussian.

3.2.2 Inference Algorithm

Assuming that maximum likelihood estimates for the component densities p_τ can be obtained, the model overall mixture model $p\left(x|\theta\right) = \sum_{\tau=1}^{K} c_\tau p_\tau(x)$ can be approximated by means of the EM algorithm. It will be convenient in the following to again represent cluster assignments in the form of binary indicator vectors z_i, with $z_{i\tau} = 1$ if site i is assigned to cluster τ and $z_{i\tau} = 0$ otherwise. The EM target function then has the form

$$Q(\theta, \tilde{\theta}) = \sum_{i,\tau} \mathbb{E}_{Z|\mathbf{x},\tilde{\theta}}\left[Z_{i\tau}\right] \log\left(c_\tau p_\tau\left(x_i|\theta_\tau\right)\right) \ . \tag{3.2.2}$$

Histogram Data Under a Parametric Model

We assume our input data to be a set $\mathbf{h} = (\mathbf{h}_1, \ldots, \mathbf{h}_n)$ of histograms, drawn i.i.d. from a source modeled by a parametric density of the form (3.2.1). Denote by I_j the interval in the data domain corresponding to bin j. For a histogram drawn from cluster τ, the probability of a data value to fall into bin I_j is $p_j^\tau\left(\Theta\right) = \int_{I_j} p_\tau\left(x|\theta_\tau\right) dx$. Given the probabilities of occurrence $p_1^\tau(\Theta), \ldots, p_{N_{\text{bins}}}^\tau(\Theta)$, the probability for any one histogram $\mathbf{h}_i = (h_{i1}, \ldots, h_{iN_{\text{bins}}})$ to occur is multinomially distributed according to

$$p_\tau(\mathbf{h}_i|\theta) := \frac{N_{\text{counts}}!}{\prod_j h_{ij}!} \prod_{j=1}^{N_{\text{bins}}} p_j^\tau(\Theta)^{h_{ij}} \ . \tag{3.2.3}$$

Including assignment variables for the EM algorithm, $\mathbf{h}$ and $\mathbf{Z}$ are jointly distributed according to

$$p(\mathbf{h}, \mathbf{z}|\theta) := \prod_i \sum_\tau z_{i\tau} c_\tau p_\tau(\mathbf{h}_i|\theta_\tau) \ . \tag{3.2.4}$$

The resulting log-likelihood is

$$\begin{aligned} l(\theta) = \sum_{i=1}^{n} \Big(\log(N_{\text{counts}}!) - \sum_{j=1}^{N_{\text{bins}}} \log(h_{ij}!) \Big) &+ \sum_{i,\tau} z_{i\tau} \log(c_\tau) \\ &+ \sum_{i,j,\tau} z_{i\tau} h_{ij} \log(p_j(\theta_\tau)) \ , \end{aligned}$$

using the standard EM trick of drawing a sum over normalized binary assignments through the logarithm. (Equality holds only because the assignment indicator vector are binary. If assignment probabilities are substituted as in the EM algorithm, the equality turns into a Jensen-type inequality, corresponding to the EM majorization inequality (2.6.13).) Since the first sum in the log-likelihood is a constant of the input data, we may drop it for the EM target function:

$$Q(\theta, \tilde{\theta}) := \sum_{i,\tau} \mathbb{E}[Z_{i\tau}] \Big(\log c_\tau + \sum_j h_{ij} \log(p_j(\theta_\tau)) \Big) \tag{3.2.5}$$

Nested EM Algorithm for the Hierarchical Model

Each component of our model (3.2.1) is again a Gaussian mixture. Optimization of the model requires a ML estimation for a simple Gaussian mixture model in the M-step. Therefore, we perform the M-step by executing an EM algorithm for each component mixture model. The approach requires hierarchical assignments: Variables for the outer EM loop, which indicate cluster assignments and will again be denoted $Z_{i\tau}$, and a complete set of assignment variables for each inner mixture, denoted $Z_{i\alpha}^\tau$, where i indicates the site, τ the cluster and α the Gaussian mode. Additionally, for the inner EM algorithm, we drop the assumption that each site is assigned to a model component: A site not assigned to the cluster in question

($z_{i\tau} = 0$ for the current cluster τ) should not be taken into account by the inner loop. Thus, $z^\tau_{i\alpha} = 1$ indicates that site i is assigned to component α of cluster τ iff $z_{i\tau} = 1$. We define effective inner assignment indicators $L^\tau_{i\alpha}$ by

$$L^\tau_{i\alpha} := Z^\tau_{i\alpha} \cdot Z_{i\tau} \ . \tag{3.2.6}$$

To make the algorithmic treatment feasible, we assume statistical independence of $Z^\tau_{i\alpha}$ and $Z_{i\tau}$. The outer algorithm computes expectations in the E-step according to

$$\mathbb{E}\left[Z_{i\tau}\right] = \frac{c_\tau p_\tau\left(\mathbf{h}_i|\theta_\tau\right)}{\sum_\nu c_\nu p_\nu\left(\mathbf{h}_i|\theta_\nu\right)} \ . \tag{3.2.7}$$

The M-step computes the mixture weights c_τ from the outer assignments as $c_\tau = \sum_i \mathbb{E}\left[z_{i\tau}\right]/n$. The inner loop consists of one EM algorithm for each cluster, which is initialized by the final inner model parameters obtained for the current cluster by previous execution of the inner loop (i. e. during the previous step of the outer algorithm). The E-step computes expectations as

$$\mathbb{E}\left[Z^\tau_{i\alpha}\right] = \frac{c^\tau_\alpha p^\tau_\alpha\left(\mathbf{h}_i|\theta^\tau_\alpha\right)}{\sum_\nu c^\tau_\nu p^\tau_\nu\left(\mathbf{h}_i|\theta^\tau_\nu\right)} \ , \tag{3.2.8}$$

where, in our case, $p^\tau_\alpha\left(\mathbf{h}_i|\theta^\tau_\alpha\right) = g\left(\mathbf{h}_i|\mu^\tau_\alpha, \sigma^\tau_\alpha\right)$. Since we assume independence, we can compute expectations for the effective inner assignments $L^\tau_{i\alpha}$ as

$$\mathbb{E}\left[L^\tau_{i\alpha}\right] = \mathbb{E}\left[Z^\tau_{i\alpha}\right] \cdot \mathbb{E}\left[Z_{i\tau}\right] \ . \tag{3.2.9}$$

The M-steps require one target function for each cluster:

$$Q^\tau(\theta^\tau, \tilde{\theta}^\tau) = \sum_{i,\alpha} \mathbb{E}\left[L^\tau_{i\alpha}\right] \log\left(c^\tau_\alpha p^\tau_\alpha\left(\mathbf{h}_i|\theta^\tau_\alpha\right)\right) \ . \tag{3.2.10}$$

By substituting histogram probabilities as in (3.2.5), we obtain

$$Q^\tau(\theta^\tau, \tilde{\theta}^\tau) = \sum_i J\left(\mathbf{h}_i\right) + \sum_{i,\alpha} \mathbb{E}\left[L^\tau_{i\alpha}\right] \log\left(c^\tau_\alpha\right) + \sum_{i,\alpha,j} h_{ij} \mathbb{E}\left[L^\tau_{i\alpha}\right] \log\left(p^\tau_{\alpha j}\left(\theta^\tau_\alpha\right)\right) , \tag{3.2.11}$$

where $J\left(\mathbf{h}_i\right)$ denotes the constant term depending only on the input data, which can again be neglected for optimization purposes. Of the two remaining terms, one depends only on the inner mixture weights c^τ_α and one on the mode parameters θ^τ_α. Therefore, the two terms can be optimized independently. Solving for the mixture weights gives

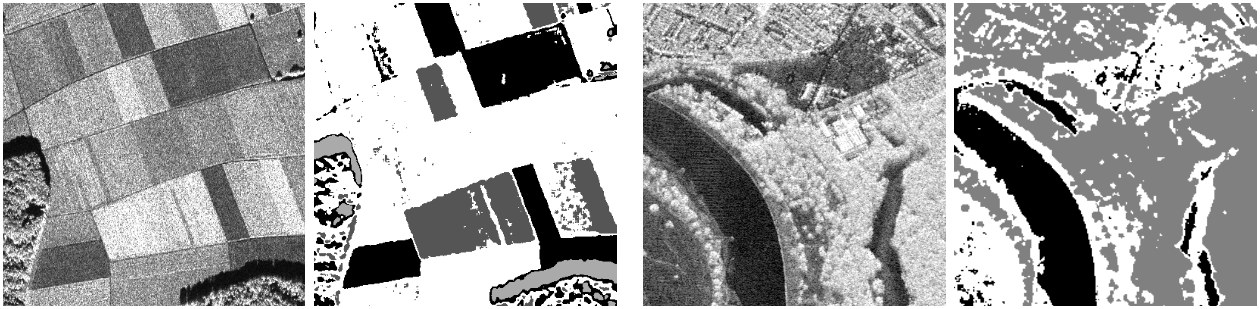
$$c^\tau_\alpha := \frac{\sum_i \mathbb{E}\left[L^\tau_{i\alpha}\right]}{\sum_{i,\alpha} \mathbb{E}\left[L^\tau_{i\alpha}\right]} = \frac{\sum_i \mathbb{E}\left[L^\tau_{i\alpha}\right]}{c_\tau} \ . \tag{3.2.12}$$

ML estimation for the Gaussian parameters during the inner M-step has to be conducted by numerical optimization of the last term in (3.2.11), because ML equations for rectified Gaussians lack closed-form solutions. The last term of (3.2.11) may be regarded, up to histogram normalization, as a cross-entropy between the average cluster data distribution and the discretized cluster model distribution. This can be turned into the negative Kullback-Leibler divergence between the two discrete distributions by adding the average data distribution's entropy (Cover and Thomas, 1991). ML estimation is therefore equivalent to minimization of the KL divergence between the data and the discretized model. Instead of computing ML estimators for the rectified model, we minimize the KL divergence on the restricted domain. The algorithm is an example of a hierarchical (recursive) application of the block optimization strategy (cf. Sec. 2.6). The outer loop is a two-block scheme. One of the blocks (the M-step) is again solved by blocking variables according to the cluster structure.

As a stopping criterion for both the outer and inner EM algorithm, we can threshold the change in assignments between consecutive steps. During the first steps of the algorithm, however, the assignments in the outer loop are still subject to large changes. It turns out that, by gradually increasing the number of inner iterations with each outer step, we can obtain results comparable (and sometimes superior) to a thresholding approach. The outer loop can then be interpreted as a generalized EM algorithm (McLachlan and Krishnan, 1997), since the M-step (the inner EM loop) is not designed to fully maximize the log-likelihood, only to increase it.

3.2.3 Application to SAR data

SAR image segmentation is an interesting application for mixture-of-mixtures models, because SAR data is known to be distributed in a characteristic fashion. The gamma distribution (and several other, closely related distributions) have been suggested as parametric models for this data (Oliver and Quegan, 1998). A gamma distribution can be approximated roughly by a single Gaussian, but very closely by a mixture of two or more Gaussians. For certain parameter configurations, gamma distributions are monotonically decreasing rather than peaked; these cases can be closely approximated by the right tail of a Gaussian when using a rectified model. If we assume that each segment is roughly gamma distributed, we can thus apply our algorithm to SAR image segmentation by clustering local histograms extracted from a SAR image using Gaussian mixtures. Figs. 3.2(a) and 3.2(b) show segmentation solutions obtained by our algorithm on two different SAR images. The locally correlated structure of the errors is typical for histogram



(a) 4 clusters and 3 modes per cluster.

(b) 3 clusters, 3 modes per clusters.

Figure 3.2: SAR images and segmentation solutions.

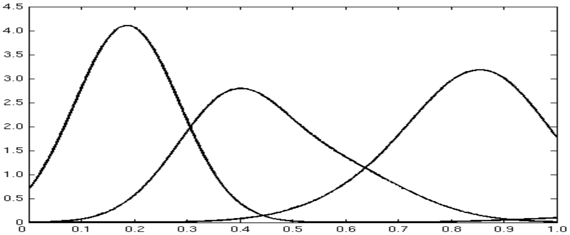


Figure 3.3: Cluster distributions (i. e. summed inner Gaussian mixtures) for the clustering solution in Fig. 3.2(b). Two modes closely resemble Gaussians, one is clearly non-Gaussian.

data with overlapping windows: Due to the size of the histogram window, local deviations from the average distribution of the segment enter in all histograms within a certain neighborhood, which are then erroneously assigned. Fig. 3.3 provides a plot of the summed Gaussian mixtures modeling the three clusters for the solution in Fig. 3.2(b). The middle mode shows how a Gaussian mixture can model a distribution of typical gamma shape.

SAR image data is often processed by logarithmic transforms. Deriving appropriate model distributions for this processed data has proven rather difficult (see, for example Xie *et al.*, 2002). The shape of the resulting distribution is roughly of reversed gamma shape, i. e. it resembles a gamma distribution of inverse skewness. The Gaussian mixtures of our model are just as suited to model this kind of data as to model gamma distributions, since a Gaussian mixture approximating a gamma distribution can be turned into a distribution of reversed gamma shape by simply shifting components.

3.2.4 Discussion

The algorithm is observed to converge despite the large number of hidden variables. In general, the performance of EM algorithms is known to deteriorate as the number of hidden variables (and thus their total entropy) increases. For the algorithm introduced above, each of the inner EM algorithms works only on a fixed subset of hidden variables, leaving all others untouched; therefore, only the entropy of the hidden variables in the subset is relevant for the performance of the corresponding inner loop. Since these subsets are pairwise disjoint, sequential execution of the inner algorithms performs a consecutive series of steps on orthogonal subspaces of the space spanned by the hidden variables, as a refinement of the alternating series of orthogonal maximization steps performed by a standard EM algorithm. In terms of Sec. 2.6, the algorithm employs an adaptive blocking approach on both levels of the model hierarchy. An inner EM loop is called in a recursive manner for each block of variables identified by the outer algorithm.

In theory, the hierarchical structure of both the model and our algorithm could be extended to a nesting depth greater than two, but due to the increasingly complicated structure of the hierarchy of hidden variables and the question of model identifiability, optimization of models with more than two layers is unlikely to be reliable. The image segmentation approach was described here only for grayscale or single-channel image data. It can easily be extended to multiple channels (without increasing the number of hidden variables) by using marginal histograms for each channel. In this case, the EM target function becomes a simple sum over the channels.

Multiple channels will be discussed in more detail in Sec. 4.1.5. A possible variation of the algorithm may be along the lines of the generalized EM (GEM) algorithm. Since short inner iterations seem to be advisable during the first few loops of the overall algorithm, one might consider a stopping criterion for the inner loop depending on the change of assignments in the outer loop, so optimization in the inner loop becomes increasingly precise as cluster assignments become more reliable.

Chapter 4

Clustering: Nonparametric Mixtures

The clustering models described in the previous chapter are standard finite mixture models, albeit with somewhat non-standard component distributions. The models described in this section are of nonparametric Bayesian type, in the sense that both are built around a mixture model in which the mixing distribution is drawn from a Dirichlet process. The first model (Sec. 4.1) constrains the clustering solution of a Dirichlet process mixture to be spatially smooth, by means of a Markov random field, and is applied to the segmentation of noisy images. This addresses the problem of *model order selection* on a given instance. The second model (Sec. 4.2), motivated by the video segmentation problem, in which a clustering problem has to be solved repeatedly along a time series, performs *model order adaptation.* The number of clusters is adjusted over time, as new components appear in or vanish from the data.

Acknowledgment. *All research presented in this chapter was conducted under supervision of Joachim M. Buhmann. The experimental results presented in Sec. 4.2 are due to experiments conducted by Samuel L. Braendle as part of a term project. Samuel made the algorithm applicable to real-world videos by significantly improving the feature extraction.*

4.1 Smoothness-Constrained Model Order Selection

Statistical approaches to image segmentation usually differ in two difficult design decisions, i. e. the statistical model for an individual segment and the number of segments which are found in the image. k-means clustering of gray or color values (Samson *et al.*, 2000), histogram clustering (Puzicha *et al.*, 1999) or mixtures of Gaussians (Hermes *et al.*, 2002) are a few examples of different model choices. Graph theoretic methods like normalized cut or pairwise clustering in principle also belong to this class of methods, since these techniques implement versions of weighted or unweighted k-means in kernel space (Bach and Jordan, 2004; Roth *et al.*, 2003). The number of clusters poses a model order selection problem with various solutions available in the literature. Most clustering algorithms require the data analyst to specify the number of classes, based either on a priori knowledge or educated guessing. More advanced methods include strategies for automatic model order selection, i. e. the number of classes is estimated from the input data. Available model selection methods for data clustering include approaches based on coding theory and minimum description length (Rissanen, 1983), and cross-validation approaches, such as stability (Lange *et al.*, 2004).

We consider a nonparametric Bayesian approach based on Dirichlet process mixture models The number of clusters as an input constant is substituted by a random variable with a control parameter. Instead of specifying a constant number of clusters, the user specifies a level of cluster resolution by adjusting the parameter. These models have been applied to a variety of problems in statistics and language processing; to the best of our knowledge, their application to image segmentation has not yet been studied. Using DPM models for segmentation seems appealing, since these models generalize parametric mixture models, and therefore one of the standard modeling tools used in data clustering. Feature extraction and grouping approaches used with finite mixture models can be transferred to the DPM framework in a straightforward way.

A possible weakness of clustering approaches for image segmentation is related to their lack of spatial structure, i. e. these models neglect the spatial smoothness of natural images: Image segmentation is performed by grouping local features (such as local intensity histograms), and only information implicit in these features is exploited in the search for satisfactory segmentations. Noisy images with unreliable features may result in incoherent segments and jagged boundaries. This drawback can be addressed by introducing spatial coupling between adjacent image features. The classic Bayesian approach to spatial statistical modeling is based on Markov random field priors (Sec. 2.7). It is widely applied in image processing to problems like image restauration and texture segmentation. As will be shown below, MRF priors which model spatial smoothness constraints on cluster labels can be combined with DPM clustering models in a natural manner. Both models are Bayesian by nature, and inference of the combined model may be conducted by Gibbs sampling.

4.1.1 Setting and Notation

The derivations of the model and inference algorithm in the following two sections are rather general, but it may be helpful briefly sketch the image segmentation problem first, to serve as illustration for the quantities arising in computations below. Image data consists of vector-valued quantities $\mathbf{x}_i$ measured at each image site ($i = 1, \dots, n$ indexes sites). A *site* is a location in the image. We will always assume sites to be arranged in a equidistant, rectangular grid within the image. In the simplest case, each pixel forms a site, but modern computer vision methods are striving more and more to abstract from pixels as reference units of image resolution, since available digital images may differ in pixel resolution by several orders of magnitude. The measurement vector $\mathbf{x}_i$ may be a grayscale value ($\mathbf{x}_i \in \mathbb{R}$), color information ($\mathbf{x}_i \in \mathbb{R}^3$), a histogram ($\mathbf{x}_i \in \mathrm{Sim}\,(\mathbb{R}, N_{\mathrm{bins}})$), or contain any other localized image information, such as saturation, hue, or filter response values. For segmentation, we assume that the image is subdividable into K groups of pixels, the *segments*, and that each segment is characterized by a different distribution of the measurements $\mathbf{x}$. Segments are spatially separated and hence mutually exclusive. Therefore, if the segment distributions are completely characterizing the pixel information contained in x, in the sense that x is conditionally independent given the segment information, a finite-size is a finite mixture of the segment distributions. By means of the Dirichlet process, K is modeled as a random variable. We represent each segment by a parametric family model $F(\,.\,|\theta_k^*)$, and make two assumptions: One is that all such F are exponential family models. The second is that the models for all segments belong to the same family of models, and differ only in the value of the parameter θ_k (though different types of models may be combined to form F, to account for heterogeneous types of measurements collected in the vector $\mathbf{x}$). To solve the segmentation problem, we estimate values of the parameters θ_k^* for each segment, and assignments of pixels to segments. The assignments will be denoted $Z_i \in \{1, \dots, K\}$, and $Z_i = k$ encodes assignment of image site i to cluster k. In the derivation below, we will also use the notation θ_i for the parameter value (and hence the segment information) at site i. If assignments Z_i are known, this is a shorthand for $\theta_i = \theta^*_{Z_i}$. By regarding the image sites as nodes of a grid, and adding edges between adjacent nodes, we obtain a neighborhood graph for a Markov random field. The parameters θ_i at each site are generated by means of a Dirichlet process prior, and we will discuss in the following how the prior can be constrained by MRF coupling between neighboring sites, so as to favor spatially smooth segmentation solutions.

4.1.2 Dirichlet process mixtures constrained by Markov random fields

Spatially constrained Dirichlet process mixture models are composed of a MRF term for spatial smoothness and a site specific data term. This local data term is drawn from a DP, whereas the interaction term may be modeled by a cost function. We will demonstrate that the resulting model defines a valid MRF. Provided that the full conditionals of the MRF interaction term can be efficiently evaluated, the full conditionals of the constrained DPM/MRF model can be efficiently evaluated as well.

The MRF distribution Π may be decomposed into a sitewise term P and the remaining interaction term M. In the cost function (2.7.4), sitewise terms correspond to singleton cliques $C = \{i\}$, and interaction terms to cliques of size two or larger. We denote the latter by $\mathcal{C}_2 := \{C \in \mathcal{C} | |C| \geq 2\}$. The MRF distribution is rewritten as

$$\begin{aligned}
\Pi(\theta_1, \ldots, \theta_n) &\propto P(\theta_1, \ldots, \theta_n) G_{\text{MRF}}(\theta_1, \ldots, \theta_n) \quad \text{with} \\
P(\theta_1, \ldots, \theta_n) &:= \frac{1}{Z_P} \exp\Big(-\sum_i H_i(\theta_i)\Big) \\
G_{\text{MRF}}(\theta_1, \ldots, \theta_n) &:= \frac{1}{Z_{\text{MRF}}} \exp\Big(-\sum_{C \in \mathcal{C}_2} H_C(\theta_C)\Big) \,.
\end{aligned} \tag{4.1.1}$$

To construct a MRF-constrained Dirichlet process prior, the marginal distribution $m_{\text{Z}}(\theta_i)$ of θ_i at each site is defined by a single random draw from a DP. The generative representation of the resulting model is

$$\begin{aligned}
(\theta_1, \ldots, \theta_n) &\sim G_{\text{MRF}}(\theta_1, \ldots, \theta_n) \prod_{i=1}^{n} m_{\text{Z}}(\theta_i) \\
m_{\text{Z}} &\sim \text{DP}\left(\alpha G_0\right) \,.
\end{aligned} \tag{4.1.2}$$

The component P in (4.1.1), defined in terms of the cost function $H_i(\theta_i)$, has thus been replaced by a random $m_{\text{Z}} \sim \text{DP}\left(\alpha G_0\right)$. To formally justify this step, we may assume a draw m_{Z} to be given and define a cost function for individual sites in terms of m_{Z}:

$$\begin{aligned} H_i(\theta_i) &:= -\log m_{\mathrm{Z}}(\theta_i) && (4.1.3)\\ Z_m &:= \int \prod_{i=1}^{n} \exp(-\log m_{\mathrm{Z}}(\theta_i)) d\theta_1 \cdots d\theta_n && (4.1.4) \end{aligned}$$

Since the term acts only on individual random variables, substitution into the MRF will not violate the conditions of the Hammersley-Clifford theorem. When the parameters $(\theta_1, \dots, \theta_n)$ are drawn from $m_{\mathrm{Z}} \sim \mathrm{DP}\,(\alpha G_0)$, the θ_i are conditionally independent given m_{Z} and their joint distribution assumes the product form

$$P(\theta_1, \dots, \theta_n | m_{\mathrm{Z}}) = \prod_{i=1}^{n} m_{\mathrm{Z}}(\theta_i)\ . \qquad (4.1.5)$$

This conditional independence of θ_i justifies the product representation (4.1.2). The model is combined with a parametric likelihood $F(\,.\,|\theta)$ by assuming the observed data $\mathbf{x}_1, \dots, \mathbf{x}_n$ to be generated according to

$$\begin{aligned} (\mathbf{x}_1, \dots, \mathbf{x}_n) &\sim \prod_{i=1}^{n} F(\mathbf{x}_i|\theta_i)\\ (\theta_1, \dots, \theta_n) &\sim G_{\mathrm{MRF}}(\theta_1, \dots, \theta_n) \prod_{i=1}^{n} m_{\mathrm{Z}}(\theta_i)\\ m_{\mathrm{Z}} &\sim \mathrm{DP}\,(\alpha G_0)\ . && (4.1.6) \end{aligned}$$

Full conditionals $\Pi(\theta_i|\theta_{-i})$ of the model can be obtained up to a constant as a product of the full conditionals of the components:

$$\Pi(\theta_i|\theta_{-i}) \propto P(\theta_i|\theta_{-i}) G_{\mathrm{MRF}}(\theta_i|\theta_{-i}) \qquad (4.1.7)$$

For DP models, $P(\theta_i|\theta_{-i})$ is computed from (4.1.5), by conditioning on θ_{-i} and integrating out the randomly drawn distribution m_{Z}. The resulting conditional prior is

$$P\left(\theta_i|\theta_{-i}\right) = \sum_{k=1}^{K} \frac{n_k^{-i}}{n-1+\alpha} \delta_{\theta_k^*}\left(\theta_i\right) + \frac{\alpha}{n-1+\alpha} G_0\left(\theta_i\right)\ . \qquad (4.1.8)$$

n_k^{-i} denotes the number of samples in group k, with the additional superscript indicating the exclusion of θ_i. The θ_i are now statistically dependent after m_{Z} is integrated out of the model. The constrained model exhibits the key property that the MRF interaction term does not affect the base measure term G_0 of the DP prior. More formally, $G_{\text{MRF}}(\theta_i|\theta_{-i})G_0$ is equivalent to G_0 almost everywhere, i. e. everywhere on the infinite domain except for a finite set of points. The properties of G_0 are not changed by its values on a finite set of points for operations such as sampling or integration against non-degenerate functions. Since sampling and integration are the two modes in which priors are applied in Bayesian inference, all computations involving the base measure are significantly simplified. Sec. 4.1.3 will introduce a sampling algorithm based on this property.

Assume that $G_{\text{MRF}}(\theta_i|\theta_{-i})$ is the full conditional of an MRF interaction term, with a cost function satisfying (2.7.8). Combining P with M yields

$$G_{\text{MRF}}(\theta_i|\theta_{-i})P(\theta_i|\theta_{-i}) \propto G_{\text{MRF}}(\theta_i|\theta_{-i})\sum_k n_k^{-i}\delta_{\theta_k^*}(\theta_i)+\alpha G_{\text{MRF}}(\theta_i|\theta_{-i})G_0(\theta_i) \; .$$

As an immediate consequence of the cost function property (2.7.8), the support of H is at most the set of the cluster parameters $\Theta^* := \{\theta_1^*, \dots, \theta_K^*\}$,

$$\operatorname{supp}(H(\theta_i|\theta_{-i})) \subset \theta_{-i} \subset \Theta^* \; . \tag{4.1.9}$$

Since Θ^* is a finite subset of the infinite domain Ω_θ of the base measure, $G_0(\Theta^*) = 0$. A random draw from G_0 will not be in Θ^* with probability 1, and hence $\exp(-H(\theta_i|\theta_{-i})) = 1$ almost surely for $\theta_i \sim G_0(\theta_i)$. With $G_{\text{MRF}}(\theta_i|\theta_{-i}) = \frac{1}{Z_{\text{MRF}}}$ almost surely,

$$G_{\text{MRF}}(\theta_i|\theta_{-i})G_0(\theta_i) = \frac{1}{Z_{\text{MRF}}}G_0(\theta_i) \tag{4.1.10}$$

almost everywhere. Sampling $G_{\text{MRF}}(\theta_i|\theta_{-i})G_0(\theta_i)$ is therefore equivalent to sampling G_0. Integration of $G_{\text{MRF}}(\theta_i|\theta_{-i})G_0(\theta_i)$ against a non-degenerate function f yields

$$\begin{aligned}\int_{\Omega_\theta} f(\theta_i)G_{\text{MRF}}(\theta_i|\theta_{-i})G(\theta_i)d\theta_i &= \int_{\Omega_\theta} \frac{f(\theta_i)}{Z_{\text{MRF}}}\exp(-H(\theta_i|\theta_{-i}))G_0(\theta_i)d\theta_i \\ &= \int_{\Omega_\theta\setminus\Theta^*} f(\theta_i)\frac{1}{Z_{\text{MRF}}}\exp(-H(\theta_i|\theta_{-i}))G_0(\theta_i)d\theta_i \\ &= \frac{1}{Z_{\text{MRF}}}\int_{\Omega_\theta} f(\theta_i)G_0(\theta_i)d\theta_i \; .\end{aligned} \tag{4.1.11}$$

The MRF constraints change only the finite component of the DPM model (the weighted sum of Dirac measures), and the full conditional of Π almost everywhere assumes the form

$$\Pi(\theta_i|\theta_{-i}) \propto \sum_{k=1}^{K} G_{\text{MRF}}(\theta_i|\theta_{-i})n_k^{-i}\delta_{\theta_k^*}(\theta_i) + \frac{\alpha}{Z_{\text{MRF}}}G_0(\theta_i) \; . \tag{4.1.12}$$

The formal argument above permits an intuitive interpretation: The finite component represents clusters already created by the model. The smoothness constraints on cluster assignments model a local consistency requirement: consistent assignments are encouraged within neighborhoods. Therefore, the MRF term favors two adjacent sites to be assigned to the same cluster. Unless the base measure G_0 is finite, the class parameter drawn from G_0 will differ from the parameters of all existing classes with probability one. In other words, a draw from the base measure always defines a new class, and the corresponding site will not be affected by the smoothness constraint, as indicated by equation (4.1.10).

4.1.3 Sampling

Application of the constrained DPM model requires a method to estimate a state of the model from data. Inference for DPM and MRF models is usually handled by Markov chain Monte Carlo sampling. Since full conditionals of sufficiently simple form are available for both models, Gibbs sampling in particular is applicable. We propose a Gibbs sampler for estimation of the combined DPM/MRF model, based on the full conditionals derived in the previous section.

A sampler for the DPM/MRF model can be obtained by adapting MacEachern's algorithm (Sec. 2.6.3) to the full conditionals of the constrained model, which were computed in the previous section. We define the algorithm before detailing its derivation. Let G_0 be an infinite probability measure, i. e. a non-degenerate measure on an infinite domain Ω_θ. Let F be a likelihood function such that F, G_0 form a conjugate pair. Assume that G_0 can be sampled by an efficient algorithm. Let H be a cost function of the form (2.7.8), and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ a set of observations drawn from the nodes of the MRF. Then the DPM/MRF model (4.1.6) can be sampled by the following procedure:

Initialize: Generate a single cluster containing all points:

$$\theta_1^* \sim G_0\left(\theta_1^*\right) \prod_{i=1}^{n} F\left(\mathbf{x}_i | \theta_1^*\right) \ . \tag{4.1.13}$$

Repeat:

1. Generate a random permutation σ of the data indices.
2. *Assignment step.* For $i = \sigma(1), \ldots, \sigma(n)$:

   (a) If $\mathbf{x}_i$ is the only observation assigned to its cluster $k = Z_i$, remove this cluster.

(b) Compute the cluster probabilities

$$\begin{aligned} q_{i0} &\propto \alpha \int_{\Omega_\theta} F\left(\mathbf{x}_i|\theta\right) G_0\left(\theta\right) d\theta \qquad (4.1.14) \\ q_{ik} &\propto n_k^{-i} \exp\left(-H(\theta_k^*|\theta_{-i})\right) F\left(\mathbf{x}_i|\theta_k^*\right) \end{aligned}$$

for $k = 1, \ldots, K$.

(c) Draw random index k according to finite distribution $(q_{i0}, \ldots, q_{iK})$.

(d) Assignment:

- If $k \in \{1, \ldots, K\}$, assign $\mathbf{x}_i$ to cluster k: Set $Z_i := k$.
- If $k = 0$, create a new cluster for $\mathbf{x}_i$:
  - Draw $\theta^*_{K+1} \sim G_0(\theta^*_{K+1})F(\mathbf{x}_i|\theta^*_{K+1})$.
  - Set $Z_i := K + 1$.

3. *Parameter update step.* For each cluster $k = 1, \ldots, K$: Update the cluster parameters θ_k^* given the class assignments $Z_1, \ldots, Z_n$ by sampling

$$\theta_k^* \sim G_0\left(\theta_k^*\right) \prod_{i|Z_i=k} F\left(\mathbf{x}_i|\theta_k^*\right) \ . \qquad (4.1.15)$$

Estimate assignment mode: For each point, choose the cluster it was assigned to most frequently during a given final number of iterations.

The sampler is implemented as a random scan Gibbs sampler, a design decision motivated by the Markov random field. Since adjacent sites couple, the data should not be scanned by index order. Initialization collects all data in a single cluster, which will result in comparatively stable results, since the initial cluster is estimated from a large amount of data. Alternatively, one may start with an empty set of clusters, such that the first cluster will be created during the first assignment step. The initial state of the model is then sampled from the single-point posterior of a randomly chosen observation, resulting in more variable estimates unless the sampler is run for a large number of iterations to ensure proper mixing of the Markov chain. The final assignment by maximization is a rather primitive form of mode estimate, but experiments show that class assignment probabilities tend to be pronounced after a sufficient number of iterations. The estimates are therefore unambiguous. If strong variations in cluster assignments during consecutive iterations are observed, maximization should be substituted by a more sophisticated approach.

The algorithm is derived by computing the assignment probabilities q_{ik} and the cluster posterior (4.1.15) based on the parametric likelihood F and the full conditional probabilities (4.1.12) of the DPM/MRF model. The posterior for a single observation $\mathbf{x}_i$ is

$$p\left(\theta_i|\theta_{-i},\mathbf{x}_i\right) \quad = \quad \frac{F(\mathbf{x}_i|\theta_i)\Pi(\theta_i|\theta_{-i})}{\int_{\Omega_\theta} F(\mathbf{x}_i|\theta)\Pi(\theta|\theta_{-i})d\theta} \ . \tag{4.1.16}$$

Substituting (4.1.12) for $\Pi(\theta_i|\theta_{-i})$ gives

$$\begin{aligned} p\left(\theta_i|\theta_{-i},\mathbf{x}_i\right) \propto F(\mathbf{x}_i|\theta_i)G_{\text{MRF}}(\theta_i|\theta_{-i})\sum_{k=1}^{K} n_k^{-i}\delta_{\theta_k^*}\left(\theta_i\right) \\ + F(\mathbf{x}_i|\theta_i)\cdot\frac{\alpha}{Z_{\text{MRF}}}G_0(\theta_i) \ . \end{aligned} \tag{4.1.17}$$

Probabilities of the individual components can be computed as their relative contributions to the mass of the overall model, i. e. by integrating each class component of the conditional (4.1.17) over Ω_θ. For each cluster $k \in \{1,\dots,K\}$ of parameters, the relevant integral measure is degenerate at θ_k^*. Integrating an arbitrary function f against the degenerate measure δ_{θ_j} "selects" the function value $f(\theta_j)$. Hence,

$$\int_{\Omega_\theta}\delta_{\theta_k^*}\left(\theta_i\right)\frac{F(\mathbf{x}_i|\theta_i)}{Z_{\text{MRF}}}\exp(-H(\theta_i|\theta_{-i}))d\theta_i = \frac{1}{Z_{\text{MRF}}}F(\mathbf{x}_i|\theta_k^*)\exp(-H(\theta_k^*|\theta_{-i})) \ . \tag{4.1.18}$$

The MRF normalization constant Z_{MRF} appears in all components and may be neglected. Combined with the coefficients of the conditional posterior (4.1.16), the class probabilities q_{i0} and q_{ij} are thus given by (4.1.14). The class posterior for sampling each cluster parameter θ_k^* is

$$\theta_k^* \sim G_0(\theta_k^*)\prod_{i|Z_i=k}F(\mathbf{x}_i|\theta_k^*)G_{\text{MRF}}(\theta_k^*|\theta_{-i}) \ . \tag{4.1.19}$$

Once again, a random draw $\theta \sim G_0$ from the base measure will not be an element of Θ^* a. s., and

$$F(\mathbf{x}_i|\theta_k^*)G_{\text{MRF}}(\theta_k^*|\theta_{-i}) = F(\mathbf{x}_i|\theta_k^*)\frac{1}{Z_{\text{MRF}}} \tag{4.1.20}$$

almost everywhere for a non-degenerate likelihood. Therefore, θ_k^* may equivalently be sampled as

$$\theta_k^* \sim G_0(\theta_k^*)\prod_{i|Z_i=k}F(\mathbf{x}_i|\theta_k^*) \ , \tag{4.1.21}$$

which accounts for the second step of the algorithm.

If F and G_0 form a conjugate pair, the integral in (4.1.14) has an analytical solution, and the class posterior (4.1.21) is an element of the same model class as G_0. If G_0 can be sampled, then the class posterior can be sampled as well. Consequently, just as MacEachern's algorithm, the algorithm above is feasible in the conjugate case. The fact that the clustering cost function gives a uniform contribution a. e. is therefore crucial. With the inclusion of the MRF contribution, the model is no longer conjugate. Due to the finite support of the cost function, however, it reduces to the conjugate case for both steps of the algorithm relying on a conjugate pair.

MacEachern's algorithm is not the only possible approach to DPM sampling. More straightforward algorithms draw samples from the posterior (4.1.16) directly, rather than employing the two-stage sampling scheme described above (Escobar, 1994). For the DPM/MRF model, the two-stage approach is chosen because of its explicit use of class labels. The choice is motivated by two reasons: First, the MRF constraints act on class assignments, which makes an algorithm operating on class labels more suitable than one operating on the parameters θ_i. The second reason similarly applies in the unconstrained case, and makes MacEachern's algorithm the method of choice for many DPM sampling problems. If a large class exists at some point during the sampling process, changing the class parameter of the complete class to a different value is possible only by pointwise updates, for each θ_i in turn. The class is temporarily separated into at least two classes during the process. Such a separation is improbable, because for similar observations, assignment to a single class is more probable then assignment to several different classes. Thus, changes in parameter values are less likely, which slows down the convergence of the Markov chain. Additionally, if a separation into different classes occurs, the resulting classes are smaller and the corresponding posteriors less concentrated, causing additional scatter. The two-stage algorithm is not affected by the problem, since parameters are sampled once for each class (rather than for each site). Given the current class assignments, the posterior does not depend on any current parameter values θ_i. The difference between the two algorithms becomes more pronounced when MRF smoothness constraints are applied. For a direct, sequential parameter sampling algorithm, constraints favoring assignment of neighbors to the same class will make separation into different classes even less probable. A two-stage sampling approach therefore seems more suited for sampling the MRF-constrained model.

4.1.4 Application to image processing

We will now discuss how the previously described and developed methods can be applied to image segmentation, both with a standard DPM approach and with a DPM/MRF model. The results derived in the previous section have not assumed any restriction on the choice of base measure G_0 and likelihood F (except for the assumption that the base measure is infinitely supported). In the following, we specialize the model by choosing specific distributions for G_0, F and the MRF term M, to define a suitable histogram clustering model for use with the DPM/MRF method.

A histogram clustering model

Our approach to image segmentation is based on histogram clustering. Given a grayscale image, local histograms are extracted as features. This feature extraction is performed on a rectangular, equidistant grid, placed within the input image. Pixels coinciding with nodes of the grid are identified with sites, indexed by $i = 1, \ldots, n$. A square histogram window is placed around each site, and a histogram $\mathbf{h}_i$ is drawn from the intensity values of all pixels within the window. The size of the window (and therefore the number N_{counts} of data values recorded in each histogram) is kept constant for the whole image, as is the number N_{bins} of histogram bins. Each histogram is described by a vector $\mathbf{h}_i = (h_{i1}, \ldots, h_{iN_{\text{bins}}})$ of non-negative integers. The histograms $\mathbf{h}_1, \ldots, \mathbf{h}_n$ are the input features of the histogram clustering algorithm. They replace the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in the previous discussion.

The parameters θ_i drawn from the DP in the DPM model are, in this context, the probabilities of the histogram bins (i. e. θ_{ij} is the probability for a value to occur in bin j of a histogram at site i). Given the probabilities of the individual bins, histograms are multinomially distributed, and the likelihood is chosen according to

$$F(\mathbf{h}_i|\theta_i) \quad = \quad N_{\text{counts}}! \prod_{j=1}^{N_{\text{bins}}} \frac{\theta_{ij}^{h_{ij}}}{h_{ij}!} = \frac{1}{Z_{\text{Mult}}(\mathbf{h}_i)} \exp\Bigl(\sum_{j=1}^{N_{\text{bins}}} h_{ij} \log(\theta_{ij}) \Bigr) \, .$$

The normalization function $Z_{\text{Mult}}(\mathbf{h}_i)$ does not depend on the value of θ_i.

The prior distribution of the parameter vectors is assumed to be conjugate, and therefore a Dirichlet distribution of dimension N_{bins}. The Dirichlet distribution (Kotz *et al.*, 2000) has two parameters β, π, where β is a positive scalar and π is a N_{bins}-dimensional probability vector. It is defined by the density

$$G_0(\theta_i|\beta\pi) = \frac{1}{Z_{\text{Dir}}(\beta\pi_j)} \exp\Big(\sum_{j=1}^{N_{\text{bins}}} (\beta\pi_j - 1)\log(\theta_{ij})\Big) \ .$$

Sampling of this model will be discussed below, along with sampling of the MRF-enhanced model.

Histogram clustering with MRF constraints

Combining the histogram clustering model with a MRF constraint requires the choice of a cost function for local smoothness. We have used the simple function

$$H(\theta_i|\theta_{-i}) = -\lambda \sum_{l\in\partial(i)} \delta_{\theta_i,\theta_l} \ . \tag{4.1.22}$$

The resulting MRF will make a larger local contribution if more neighbors of site i are assigned to the same class, thereby encouraging spatial smoothness of cluster assignments.

To sample the MRF-constrained histogram clustering model, the sampler has to be derived for the particular choice of distributions (4.1.22) and (4.1.22), which requires computation of the class probabilities q_{i0} and q_{ik} in (4.1.14) and the respective posterior (4.1.15). Since F, G_0 form a conjugate pair, their product is (up to normalization) a Dirichlet density:

$$F(\mathbf{h}_i|\theta_i)G_0(\theta_i) \propto \exp\Big(\sum_j (h_{ij} + \beta\pi_j - 1)\log(\theta_{ij})\Big) = G_0(\theta_i|\mathbf{h}_i + \beta\pi) \ . \tag{4.1.23}$$

Therefore, q_{i0} has an analytic solution in terms of partition functions:

$$\begin{aligned}\int_{\Omega_\theta} F(\mathbf{h}_i|\theta_i)G_0(\theta_i)d\theta_i &= \int_{\Omega_\theta} \frac{\exp\Big(\sum_j (h_{ij} + \beta\pi_j - 1)\log(\theta_{ij})\Big)}{Z_{\text{Mult}}(\mathbf{h}_i)Z_{\text{Dir}}(\beta\pi)} d\theta_i \\ &= \frac{Z_{\text{Dir}}(\mathbf{h}_i + \beta\pi)}{Z_{\text{Mult}}(\mathbf{h}_i)Z_{\text{Dir}}(\beta\pi)} \ .\end{aligned} \tag{4.1.24}$$

For $k = 1, \dots, K$,

$$\begin{aligned} q_{ik} &\propto n_k^{-i} \exp(-H(\theta_k^*|\theta_{-i}))F(\mathbf{h}_i|\theta_k^*) \\ &= \frac{n_k^{-i}}{Z_{\text{Mult}}(\mathbf{h}_i)} \exp\Big(\lambda \sum_{l\in\partial(i)} \delta_{\theta_i,\theta_l} + \sum_j h_{ij}\log(\theta_{kj}^*)\Big) \ . \end{aligned} \tag{4.1.25}$$

Since the multinomial partition function $Z_{\text{Mult}}(\mathbf{h}_i)$ appears in all equations, the cluster probabilities may be computed for each i by computing preliminary values

$$\begin{aligned}\tilde{q}_{i0} &:= \frac{Z_{\text{Dir}}(\mathbf{h}_i + \beta\pi)}{Z_{\text{Dir}}(\beta\pi)} \\ \tilde{q}_{ik} &:= n_k^{-i} \exp\Big(\lambda \sum_{l \in \partial(i)} \delta_{\theta_i, \theta_j} + \sum_j h_{ij} \log(\theta^*_{kj})\Big) .\end{aligned} \tag{4.1.26}$$

From these, cluster probabilities are obtained by normalization:

$$q_{ik} := \frac{\tilde{q}_{ik}}{\sum_{l=0}^{K} \tilde{q}_{il}} . \tag{4.1.27}$$

The posterior to be sampled in (4.1.21) is Dirichlet as well:

$$\begin{aligned}G_0(\theta^*_k|\beta\pi) \prod_{i|Z_i=k} F(\mathbf{x}_i|\theta^*_k) &\propto \exp\Big(\sum_j (\beta\pi_j + \sum_{i|Z_i=k} h_{ij} - 1) \log(\theta^*_{kj})\Big) \\ &\propto G_0\Big(\theta^*_k \Big| \beta\pi + \sum_{i|Z_i=k} \mathbf{h}_i\Big)\end{aligned} \tag{4.1.28}$$

Dirichlet distributions can be sampled efficiently by means of Gamma sampling; cf. for example (Devroye, 1986). Sampling of the unconstrained model may be conducted by choosing $\lambda = 0$ in the MRF cost function.

Behavior of the segmentation model

Since both the base measure and the posterior sampled in the algorithm are Dirichlet, the properties of this distribution have a strong influence on the behavior of the clustering model. Dirichlet densities are delicate to work with, since they involve a product over exponentials, and because their domain covers a multidimensional real simplex, which renders them difficult to plot or illustrate. The clustering model, however, which has been obtained by combining the Dirichlet base measure and the multinomial likelihood behaves in a manner that is intuitive to understand: Each observed histogram $\mathbf{h}_i$ is assumed to be generated by the likelihood F, which is determined at each site by the parameter θ_i. The vector θ_i lies in the N_{bins}-dimensional simplex $\text{Sim}\,(\mathbb{R}, N_{\text{bins}})\, N_{\text{bins}}$, and it can be regarded as a finite probability distribution on the histogram bins. Its distribution $G_0(\theta_i|\beta\pi)$ is parameterized by another vector $\pi \in \text{Sim}\,(\mathbb{R}, N_{\text{bins}})\, N_{\text{bins}}$, which defines the expected value of G_0. The scalar parameter β controls the scatter of the distribution:

The larger the value of β, the more tightly G_0 will concentrate around π. For $\beta\pi = (1, \dots, 1)^t$, G_0 is the uniform distribution on the simplex. Consider the posterior (4.1.28), which is a Dirichlet distribution with the scaled vector $\beta\pi$ replaced by $\beta\pi + \sum_{i|Z_i=k} \mathbf{h}_i$. By setting

$$\tilde{\beta}_k := \Big\| \beta\pi + \sum_{i|Z_i=k} \mathbf{h}_i \Big\|_1 \qquad \text{and} \qquad \tilde{\pi}_k := \frac{1}{\tilde{\beta}} \Big(\beta\pi + \sum_{i|Z_i=k} \mathbf{h}_i \Big) \;, \quad (4.1.29)$$

the posterior assumes the form $G_0(\,.\,|\tilde{\beta}_k \tilde{\pi}_k)$. For each cluster k, the expected value of the posterior is $\tilde{\pi}_k$, and its scatter is determined by $\tilde{\beta}_k$. The expected value $\tilde{\pi}_k$ is the (normalized) average of the histograms assigned to the cluster, with an additive distortion caused by the base measure parameters. The larger β, the more influence the prior will have, but generally, it has less influence if the number of histograms assigned to the cluster is large. Since $\tilde{\beta}_k$ controls the scatter and grows with the number of histograms assigned, the posterior of a large cluster will be tightly concentrated around its mean. In other words, for a very large cluster, drawing from the posterior will reproduce the cluster's normalized average with high accuracy. Therefore, larger clusters are more stable. For a smaller cluster, draws from the posterior scatter, and the additive offset $\beta\pi$ has a stronger influence.

Assignments to clusters are determined by sampling from the finite distributions $(q_{i0}, \dots, q_{iK})$, which are based on the multinomial likelihood F. For illustration, consider a non-Bayesian maximum likelihood approach for F. Such an approach would assign each histogram to the class which achieves the highest likelihood score. Multinomial likelihood maximization can be shown to be equivalent to the minimization of the Kullback-Leibler divergence between the distribution represented by the histogram and that defined by the parameter. Each histogram would thus be assigned to the "nearest" cluster, in the sense of the Kullback-Leibler divergence. The behavior of our histogram clustering model is similar, with two notable differences: The greedy assignment is replaced by a sampling approach, and the DPM model may create a new class for a given histogram, instead of assigning it to a currently existing one. The key properties of the model are not affected or altered by adding or removing the MRF constraints, except for the assignment step: The assignment probabilities computed from the basic, unconstrained model are modified by the constraint to increase the probability of a smooth assignment.

4.1.5 Extensions of the constrained model

The histogram clustering model introduced in Sec. 4.1.4 characterizes image patches by a set of intensity histograms. We will extend this concept to include additional features by modeling multiple channels and side information not contained in the features. The segmentation algorithm becomes directly applicable to multi-channel data, such as color images, multiple frequency bands in radar images, or image filter response data. For color images or multi-band radar data, the segmentation algorithm can draw on marginal intensity histograms extracted from each channel. Filter responses of local image filters can be represented as an image, and be included as additional channels. For example, texture information may be included in color image segmentation by including histograms of Gabor responses.

Including multiple channels increases the amount of data. The information provided by the different channels affects the behavior of the model by means of the likelihood F. The DPM/MRF model provides a second generic way of including additional information, by using side information to adjust the edge weights w_{ij} of the MRF neighborhood graph. The w_{ij} must not depend on the current state of the model (i. e. the values of the model variables θ_i), but they may depend on the data. The coupling strength between adjacent sites may thus be modeled conditional on the local properties of the input image.

Multiple channels

The DPM/MRF histogram clustering model introduced above represents each site by a single histogram. To model multiple channels, we again assume the marginal histograms to be generated by a multinomial likelihood F, with parameter vectors drawn from a Dirichlet distribution prior G_0. For N_{ch} separate channels, a local histogram $\mathbf{h}_i^c = (h_{i1}^c, \dots, h_{iN_{\text{bins}}}^c)$ is assumed to be drawn from each channel c at each site i. The channels are parameterized individually, so each histogram $\mathbf{h}_i^c$ is associated with a bin probability vector θ_i^c with prior probability $G_0(\theta_i^c|\beta^c\pi^c)$. The joint likelihood is assumed to factorize over channels. The resulting posterior for site i has the form

$$(\theta_i^1, \dots, \theta_i^{N_{\text{ch}}}) \sim \prod_{c=1}^{N_{\text{ch}}} F(\mathbf{h}_i^c|\theta_i^c) G_0(\theta_i^c|\beta^c\boldsymbol{\pi}^c) \ . \tag{4.1.30}$$

This generalization of the DPM/MRF clustering model (Sec. 4.1.2) only affects the base measure G_0 and the random function m_{Z} in (4.1.5), it does not alter the MRF interaction term M. Both the DPM/MRF model and the sampling algorithm remain applicable. In the sampler, only the computation of the cluster probabilities q_{ik} and the cluster posterior in (4.1.15) have to be modified. Substituting the multi-channel likelihood F

into (4.1.26) yields

$$\tilde{q}_{i0} = \prod_{l=1}^{N_{\text{ch}}} \frac{Z_{\text{Dir}}(\mathbf{h}_i^c + \beta^l \pi^l)}{Z_{\text{Dir}}(\beta^c \boldsymbol{\pi}^l) Z_{\text{Mult}}(\mathbf{h}_i^c)} \tag{4.1.31}$$

and

$$\tilde{q}_{ik} = n_{-i}^k \exp\left(-H(\theta_k^* | \theta_{-i})\right) \prod_{c=1}^{N_{\text{ch}}} F(\mathbf{h}_i^c | \theta_k^{*c}) \ . \tag{4.1.32}$$

Each site remains associated with a single assignment variable Z_i (the clustering model groups sites, rather than individual histograms). The cluster posterior (4.1.15) is

$$(\theta_i^{*1}, \ldots, \theta_i^{*N_{\text{ch}}}) \sim \prod_{c=1}^{N_{\text{ch}}} G_0\Big(\theta_i^{*c} \Big| \beta^c \boldsymbol{\pi}^c + \sum_{i|Z_i=k} \mathbf{h}_i^c\Big) \ . \tag{4.1.33}$$

This model with multiple channels assumes that local marginal histograms are obtained individually from each channel. It is not applicable to joint histograms. The advantage of marginal histograms is that, unlike joint histograms, they are not affected by the curse of dimensionality. At a constant level of discretization, the number of bins in a joint histogram grows exponentially with the number of dimensions, as opposed to linear growth for a set of marginal histograms. Marginal histograms therefore provide more robust estimates and require less complex models for their representation. Their disadvantages are (i) the loss of co-occurrence information, and (ii) the independence assumption in (4.1.30) required to obtain a feasible model. Choosing marginal histograms can be justified by observing that both problems are limited by the use of local features. Marginalization of histograms can incur a substantial loss of image information. The global marginal histograms of an RGB image, for example, are informative about the amount of red and blue occurring in the image, but not about the amount of purple. The latter requires a joint histogram. Since the segmentation algorithm relies on local features, the loss of co-occurrence information is limited: If the local marginals show the occurrence of both red and blue within a small local window, a joint histogram will not provide much additional information. Joint histograms measure co-occurrence at pixels, whereas local marginal histograms coarsen the resolution from pixels to local windows. A similar argument applies for independence: The product in (4.1.30) constitutes a local independence assumption, i. e. the marginal histograms $\mathbf{h}_i^1, \mathbf{h}_i^2, \ldots$ are assumed to be independent at site i. Histograms of two different channels at two different sites (e. g. $\mathbf{h}_i^1$ and $\mathbf{h}_l^2$)

are not independent, since they interact through the cluster parameters and MRF constraints. Local independence of channels is a more accurate assumption than global independence. Loosely speaking, given a single channel of an entire color image, guessing the image structure (and therefore significant information about the remaining channels) is usually easy. This is not the case for local image patches containing only a few pixels, since their resolution is below the scale of image structures.

Side information: Image edges

Smoothing constraints may result in unsolicited coupling effects at segment boundaries. Two sites may belong to different segments and still be caused by the smoothing term to be assigned to the same cluster. Side information on image edges can be used to improve the resolution of segment boundaries, in addition to the input features of the algorithm. Edge information is particularly useful for segmentation, since segment boundaries can be expected to coincide with an image edge. A priori we assume that two sites should not be coupled by a smoothing constraint if they are separated by an image edge. Therefore, edge information may be taken into account in the DPM/MRF model by modifying the neighborhood graph $\mathcal{N}$ of the MRF:

1. Generate an edge map using a standard edge detector.
2. If two sites i and j are separated by an image edge, set $w_{ij} = w_{ji}$ to zero.

Since the MRF constraints act only along edges of the neighborhood graph, this will eliminate coupling between the features $\mathbf{h}_i$ and $\mathbf{h}_j$. Neighborhoods in the MRF graph are usually of small, constantly bounded size ($|\partial(i)| \leq 8$ for the examples provided in the following section), such that the computational expense of this preprocessing step will be linear in the number of sites (rather than quadratic, despite the pairwise comparison).

Given an edge map, i. e. a binary matrix indicating pixels which are classified as edges by the edge detector, the algorithm has to determine whether or not a given pair of sites is separated by an edge. The method used in the experiments presented in Sec. 4.1.6 is to remove sites containing an edge pixel in their local image neighborhood from all neighborhoods in $\mathcal{N}$. A single edge is then reinserted (by setting the corresponding weight to 1), such that each site links with at least one of its neighbors. The reinserted edge is chosen in the same direction for all sites (e. g. the edge connecting the site with its left neighbor in the image). This may cause an inaccuracy of edge positions, but only on the scale of the subsampling grid. Simply removing sites completely from the graph neighborhood turns out to interact unfavorably with the model selection property of the DPM algorithm: The histogram windows of sites close to segment boundaries contain mixture distributions from two segments, which typically differ significantly from other local distributions. If coupling constraints with their neighbors are removed, these edge sites tend to be assigned clusters of their own. Edges become visible in the segmentation solution as individual segments. In other words, the approach is prone to remove constraints in regions where they are particularly relevant.

Side information: Local data disparity

Alternatively, the coupling weights w_{il} may be set according to local data disparity, an approach originally introduced in (Geman *et al.*, 1990). The idea is to define a similarity measure $d(\mathbf{x}_i, \mathbf{x}_l)$ between local data vectors and set $w_{il} := d(\mathbf{x}_i, \mathbf{x}_l)$. Substitution into the cost function (2.7.7) yields

$$G_{\text{MRF}}(\theta_i|\theta_{-i}) \propto \frac{1}{Z_{\text{MRF}}} \exp\Big(-\lambda \sum_{l\in\partial(i)} d(\mathbf{x}_i, \mathbf{x}_l)\delta_{Z_i,Z_l}\Big) \qquad (4.1.34)$$

for the MRF interaction term. The point-wise contribution P is not affected. Computing the weights from data makes the partition function $Z_{\text{MRF}} = Z_{\text{MRF}}(\lambda, \mathbf{x}_i, \mathbf{x}_{\partial(i)})$ data-dependent, but the dependence is uniform over clusters at any given site, and the partition function still cancels from the computation of assignment probabilities in the same manner as described in Sec. 4.1.3. The similarity function has to be symmetric, i. e. satisfy $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$, to ensure symmetry of the edge weights. In the case of Euclidean data $\mathbf{x}_i \in \mathbb{R}^m$, for example, d may be chosen as a regularized inverse of the Euclidean distance:

$$d(\mathbf{x}_i, \mathbf{x}_l) := \frac{1}{1 + \|\mathbf{x}_i - \mathbf{x}_l\|_2} \qquad (4.1.35)$$

The corresponding edge weight w_{il} will be 1 (maximum coupling) for identical data, and decay hyperbolically as the distance between data values increases. Histograms represent finite probability distributions (up to normalization). Hence, for histogram data, norms may be substituted by distribution divergence measures, such as the Kolmogorov-Smirnov statistic (Lehmann and Romano, 2005) or the Jensen-Shannon divergence (Lin, 1991). The Kullback-Leibler divergence and chi-square statistic (Lehmann and Romano, 2005) are not directly applicable, since neither is symmetric. The dissimilarity measure should be carefully chosen for robustness, since local dissimilarities are measured between individual data values, with no averages to temper the effect of outliers. High-order statistics, such as the

Jensen-Shannon divergence, are notoriously hard to estimate from data. For practical purposes, Euclidean norms or the Kolmogorov-Smirnov distance seem a more advisable choice.

In the Bayesian setting, the MRF interaction term is part of the prior. If the MRF cost function depends on the data, the prior function depends on both the parameter variables and the data, a role usually reserved for the likelihood. The data-dependent prior may be justified as a joint distribution on data and parameters, where the data is "fixed by observation", as outlined in (Geman *et al.*, 1990). We note the formal difference: A likelihood is a function of both data and parameter, but constitutes a density only with respect to the data. The MRF prior defined above is a density with respect to the parameter variables. The cost term in (4.1.34) measures local differences between the data vectors associated with adjacent sites. The overall model can be interpreted in terms of distances: Many parametric distributions may be regarded as exponentials of average divergences between data and parameters. Multinomial and Dirichlet distributions measure divergence between data and parameters in a Kullback-Leibler sense, and the Gaussian by an average Euclidean distance. Suppose the multinomial/Dirichlet model described in Sec. 4.1.4 is combined with the cost function (2.7.7) and edge weights $w_{il} = d(\mathbf{h}_i, \mathbf{h}_l)$. The log-posterior of each cluster is a weighted sum of divergence measures, between data and parameter variables (contributed by the likelihood F), hyperparameters and parameter variables (base measure G_0) and data at adjacent sites (MRF interaction term M). The DP hyperparameter α adjusts the sensitivity with which the DP will react to the differences measured by the parametric model by creating new classes.

4.1.6 Experimental results

The experiments presented below implement both the unconstrained DPM model and the DPM/MRF model for image segmentation. The unconstrained model is applied to natural images (from the Corel database), which are sufficiently smooth not to require spatial constraints. The DPM/MRF model is applied to synthetic aperture radar (SAR) images and magnetic resonance imaging (MRI) data, chosen for their high noise level.

Model parameters

In addition to the parameter α and the input data, the estimate of the number of clusters K depends on the parametric model used with the DPM/MRF approach. The Dirichlet process estimates the number of clusters based on disparities in the data. The disparities are measured by the

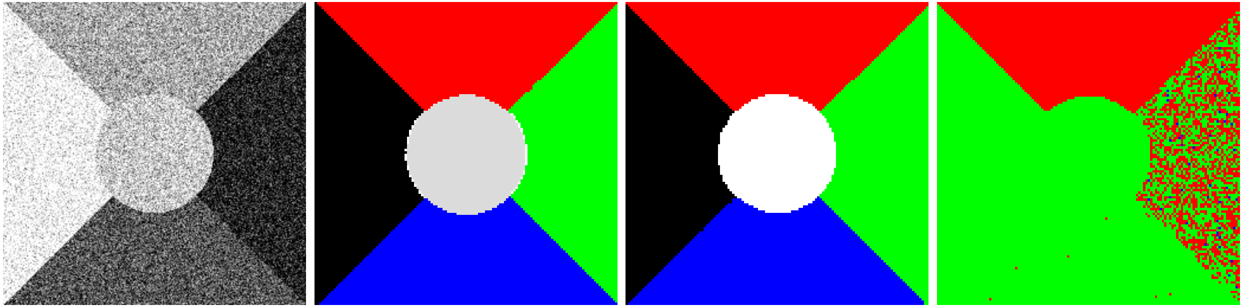


Figure 4.1: Behavior of the unconstrained DPM sampler on an image with clearly defined segments. Left to right: Input image and segmentation results for $\alpha = 10$, $\alpha = 10^{-4}$ and $\alpha = 10^{-10}$.

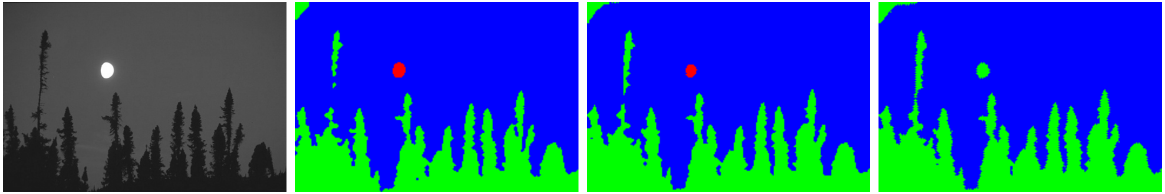


Figure 4.2: Unconstrained DPM results on a simple natural image (Corel database). Left to right: Original image, DPM results with $\alpha = 10^{-2}$, $\alpha = 10^{-7}$ and $\alpha = 10^{-9}$.

parametric model, which consists of the likelihood F and the base measure G_0. The parameters θ of F are random variables estimated during inference, but any parameters of G_0 are hyperparameters of the overall model. Adjusting the parameters of G_0 changes the parametric model, and thereby influences the model order selection results of the DP prior. In general, increasing the scatter of the distribution G_0 will increase the number of clusters in the DPM solution: The parameters θ_k^* representing the clusters are effectively sampled from a posterior with prior G_0. A concentrated distribution G_0 biases the cluster parameters towards its expected value, and restricts the adaptation of each cluster to the data it contains.

Our strategy is to set the expectation of the base measure to a generic value. The bias incurred from the base measure can then be regarded as data regularization: When a new cluster is created by the algorithm, its initial parameter is based on a single observation. The biasing effect of the hyperparameters will prevent the cluster parameter from adapting to individual outliers. As more observations are collected in the cluster, the influence of the bias decreases. The relative magnitude of the bias is determined by the scatter of the base measure.

The histogram clustering model described in Sec. 4.1.4 uses a Dirichlet distribution as its base measure, $G_0 = G_0(\,.\, | \beta\pi)$. The expected value is the parameter π and the scatter is controlled by β (increasing the value decreases scatter). Since $\pi \in \text{Sim}\,(\mathbb{R}, N_{\text{bins}})\, d$ represents a finite probability distribution, the obvious choice for a generic value is the uniform vector $\pi = (1/N_{\text{bins}}, \dots, 1/N_{\text{bins}})$, which was used for most experiments in this section. For some cases, π was chosen as the normalized average histogram of the input image, which adapts the method somewhat better to the input data than a uniform parameter vector, but also tends to result in an algo-

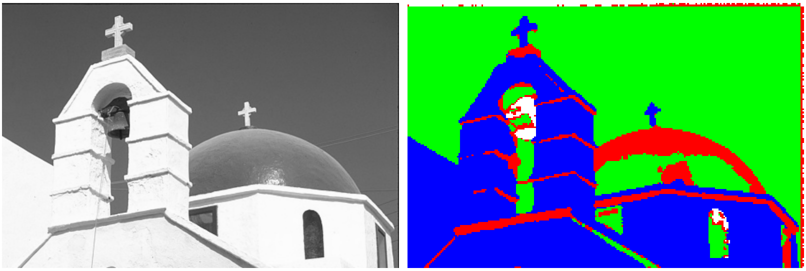


Figure 4.3: Natural image (Corel database, left) and unconstrained DPM segmentation result (right).

rithm which neglects small segments, as will be discussed below. We propose to choose a β of the same order of magnitude as the mass of a histogram ($\beta = 2N_{\text{counts}}$ was used for the experiments in this section). The regularization effect will be substantial for the creation of new clusters containing only a single histogram, and prevent overfitting of cluster representations to outliers. As soon as the cluster contains a significant number of observations (in particular when it is large enough to be visible in an image segmentation solution), the effect of the bias becomes negligible.

Image segmentation by a DPM model

As a first test of the model selection property of the DPM clustering algorithm, the (unconstrained) algorithm was applied to an image with unambiguously defined segments (the noisy Mondrian in Fig. 4.1); the classes are accurately recovered for a wide range of hyperparameter values (α ranging from 10^{-5} to 10^{1}). For a very small value of the hyperparameter ($\alpha = 10^{-10}$), the estimated number of clusters is too small, and image segments are joined erroneously.

Figs. 4.2 and 4.3 show images from the Corel database. The three classes in Fig. 4.2 are clearly discernible, and are once again correctly estimated by the process for $\alpha = 10^{-2}$ and $\alpha = 10^{-7}$. For $\alpha = 10^{-9}$, the process underestimates the number of segments. Note that this results in the deletion of the smallest segment (in this case, the moon): The scatter of the Dirichlet posterior distribution (4.1.15) is controlled by the total mass of its parameter vector ($\beta\pi + \sum_{i|Z_i=k} \mathbf{h}_i$). Since large clusters contribute more histogram mass to the parameter vector than small clusters, they are more stable (cf. Sec. 4.1.4). A small cluster defines a less concentrated posterior, and is less stable. The effect is more pronounced if π is chosen to be the average normalized histogram of the input image, since small segments will be underrepresented. If π is chosen uniform, the offset $\beta\pi$ acts as a regularization term on the average histogram.

The segmentation result in Fig. 4.3 exhibits a typical weakness of segmentation based exclusively on local histograms: The chapel roof is split into two classes, since it contains significantly different types of intensity histograms due to shading effects. Otherwise, the segmentation is precise, because the local histograms carry sufficient information about the segments.

Segmentation with smoothness constraints

The results discussed so far do not require smoothing: The presented images (Figs. 4.2 and 4.3) are sufficiently smooth, and the noise in Fig. 4.1 is

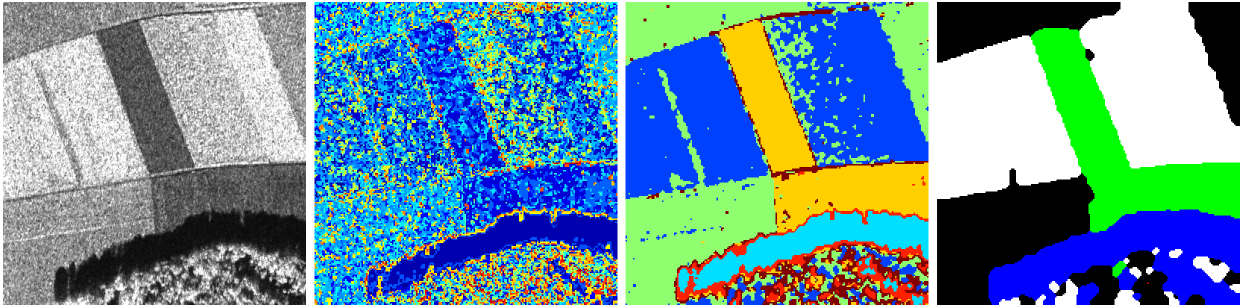


Figure 4.4: Segmentation results on real-world radar data. Left to right: Original image, unconstrained DPM segmentation, and constrained DPM segmentation at two different levels of smoothing, $\lambda = 1$ and $\lambda = 5$.

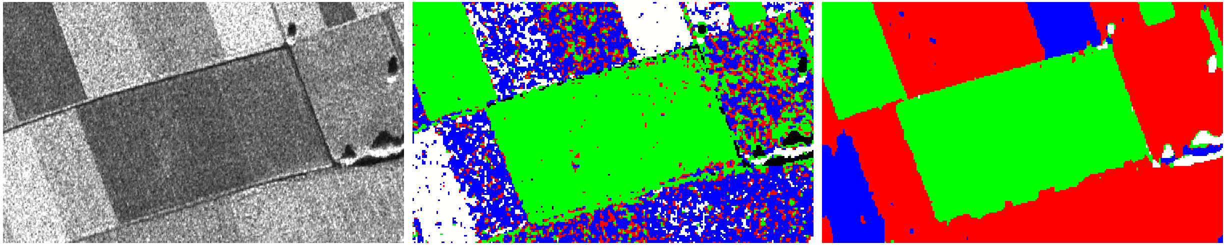


Figure 4.5: Original SAR image (left), unconstrained DPM segmentation (middle), smoothed DPM segmentation (right).

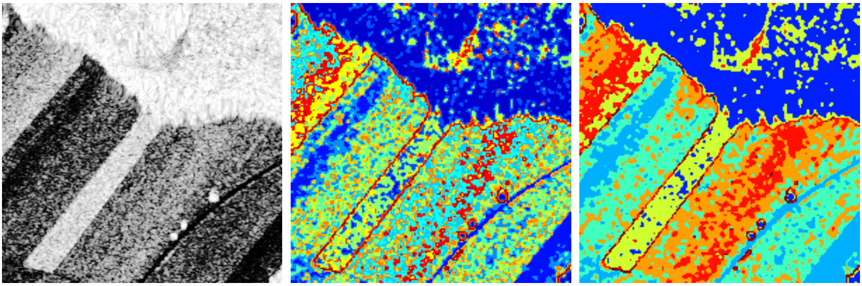


Figure 4.6: A SAR image with a high noise level and ambiguous segments (left). Solutions without (middle) and with smoothing.

additive Gaussian, which averages out well even for histograms of small image blocks.

Synthetic aperture radar (SAR) images and MRI data are more noisy than the Corel images. The images shown in Figs. 4.4 and 4.5 are SAR images of agricultural areas. In both cases, the unconstrained DPM clustering result are inhomogeneous. Results are visibly improved by the MRF smoothing constraint. Fig. 4.6 shows results for an image which is hard to segment by histogram clustering, with several smaller classes that are not well-separated and a high noise level. In this case, the improvement achievable by smoothing is limited. Results for a second common type of noisy image, MRI data, are shown in Fig. 4.8.

The Dirichlet process approach does not eliminate the class number parameter. Like any Bayesian method, it effectively replaces the parameter by a random variable, which is equipped with a prior probability. The prior is controlled by means of the hyperparameter α. The number of classes depends on α, but the influence of the hyperparameter can be overruled by observed evidence. A question of particular interest is therefore the influence of the hyperparameter α on the number of clusters. Table 4.1 shows the average number of clusters selected by the model for a wide range of hyperparameter values, ranging over several orders of magnitude. Averages are taken over ten randomly initialized experiments each. In general, the number of clusters increases monotonically with an increasing value of the DP scatter parameter α. With smoothing activated, the average estimate becomes more conservative, and more stable with respect to a changing α. The behavior of the estimate depends on the class structure of the data. If the data is well-separated, estimation results become more stable, as is the case for the MRI image (Fig. 4.8). With smoothing activated, the estimated

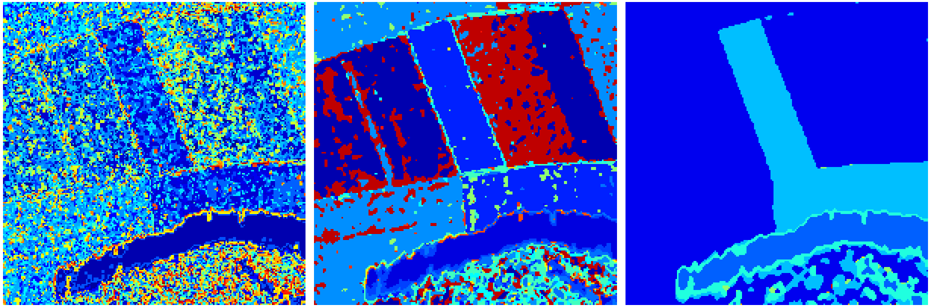


Figure 4.7: Segmentation results for $\alpha = 10$, at different levels of smoothing: Unconstrained (left), standard smoothing ($\lambda = 1$, middle) and strong smoothing ($\lambda = 5$, right).

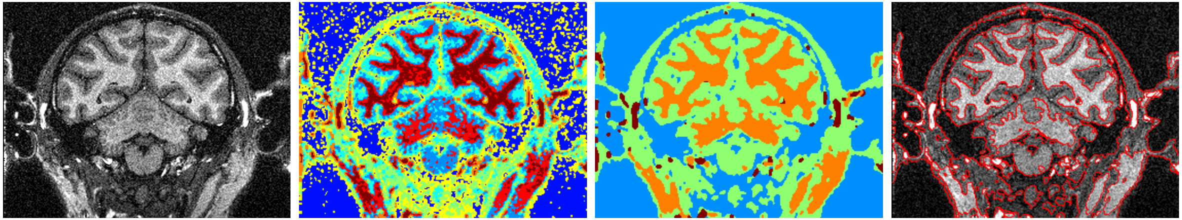


Figure 4.8: MR frontal view image of a monkey's head. Left to right: Original image, unsmoothed DPM segmentation, smoothed DPM segmentation, and original image overlaid with segment boundaries (smoothed result).

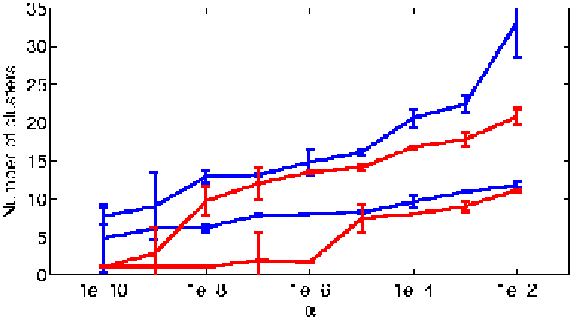


Figure 4.9: Influence of the base measure choice: Average number of clusters plotted against α, for two different values of base measure scatter. Blue curves represent $\beta = 50$, red curves $\beta = 200$. In either case, the upper curve corresponds to the unsmoothed and the lower curve to the smoothed model.

number of clusters stabilizes at $K = 4$. In contrast, the data in Fig. 4.4 does not provide sufficient evidence for a particular number of classes, and no stabilization effect is observed. We thus conclude that, maybe not surprisingly, the reliability of DPM and DPM/MRF model selection results depends on how well the parametric clustering model used with the DP is able to separate the input features into different classes. The effect of the base measure scatter, defined here by the parameter β, is demonstrated in Fig. 4.9. The number of clusters selected is plotted over α at two different values of $\beta = 50$ and $\beta = 200$, each with and without smoothing. The number of clusters is consistently decreased by increasing β and activating the smoothing constraint.

The stabilizing effect of smoothing is particularly pronounced for large values of α, resulting in a large number of clusters selected by the standard DPM model. Results in Fig. 4.7 were obtained with $\alpha = 10$, which results in an over-segmentation by the DPM model ($\bar{K} = 87.1$). With smoothing, the estimated number of clusters decreases ($\bar{K} = 29.1$). The level of smoothing can be increased by scaling the cost function. By setting $\lambda = 5$, the number of clusters is decreased further, to $\bar{K} = 8.2$.

Extensions: Edges and multiple channels

Long runs of the sampler with a large value of λ, which may be necessary on noisy images to obtain satisfactory solutions, can result in unso-

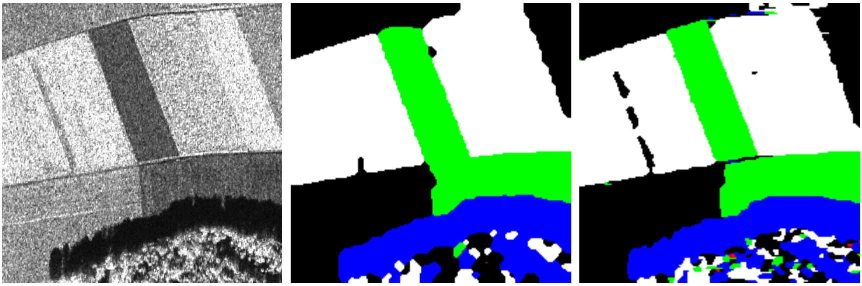


Figure 4.10: Stabilization of segmentation results by edge information for a strong smoothing constraint: Smoothed segmentation (middle), and the same experiment repeated using edge information (right), both conducted on the image in Fig. 4.4.

	Image Fig. 4.4		Image Fig. 4.8	
α	DPM	smoothed	DPM	smoothed
1e-10	7.7 ± 1.1	4.8 ± 1.4	6.3 ± 0.2	2.0 ± 0.0
1e-8	12.9 ± 0.8	6.2 ± 0.4	6.5 ± 0.3	2.6 ± 0.9
1e-6	14.8 ± 1.7	8.0 ± 0.0	8.6 ± 0.9	4.0 ± 0.0
1e-4	20.6 ± 1.2	9.6 ± 0.7	12.5 ± 0.3	4.0 ± 0.0
1e-2	33.2 ± 4.6	11.8 ± 0.4	22.4 ± 1.8	4.0 ± 0.0

Table 4.1: Average number of clusters (with standard deviations), chosen by the algorithm on two images for different values of the hyperparameter. When smoothing is activated ($\lambda = 5$, right column), the number of clusters tends to be more stable with respect to a changing α.

licited smoothing effects. Comparing the two smoothed solutions in Fig. 4.4 (lower left and right), for example, shows that a stronger smoothing constraint leads to a deterioration of some segment boundaries. The segment boundaries can be stabilized by including edge information as described in Sec. 4.1.5. An example result is shown in Fig. 4.10.

For SAR images consisting of multiple frequency bands, the multi-channel version of the DPM/MRF model (Sec. 4.1.5) can be applied. A segmentation result is shown in Fig. 4.11. Both solutions were obtained with smoothing. To demonstrate the potential value of multiple channel information, only a moderate amount of smoothing was applied. One solution (middle) was obtained by converting the multi-channel input image into a single-

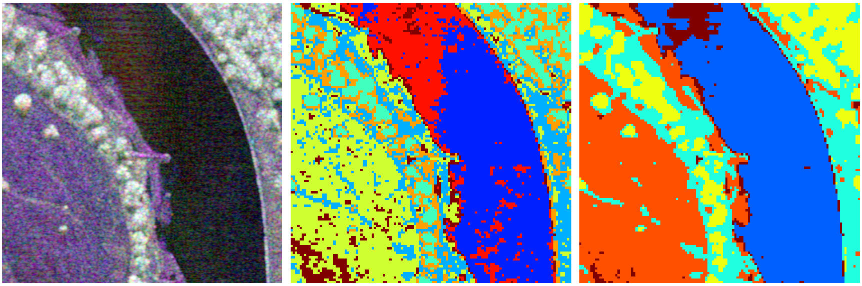


Figure 4.11: Multi-channel information: A SAR image consisting of three frequency bands (left), segmentation solutions obtained from the averaged single channel by the standard MRF/DPM model (middle) and by the multi-channel model (right).

channel grayscale image before applying the DPM/MRF model. The second solution (right) draws explicitly on all three frequency bands by the multi-channel model. Parameter values for the single-channel and multi-channel approach are not directly comparable. When computing the cluster assignment probabilities q_{ik}, the multi-channel model multiplies probabilities over channels. Hence, the computed values are generally smaller than in the single-channel case. This increases the relative influence of α, and the multi-channel approach tends to select more clusters for the same parameter values than the single-channel model. To make the result comparable, we have chosen examples with similar number of clusters ($K = 7$ and $K = 5$, respectively). The segmentation result is visibly improved by drawing on multi-channel features.

Comparison: Stability

Relating the approach to other methods is not straightforward, since model order selection methods typically try to estimate a unique, “correct” number of clusters. We use the *stability method* to devise a comparison that may offer some insight into the behavior of the DPM model.

Stability-based model selection for clustering (Dudoit and Fridlyand, 2002; Breckenridge, 1989; Lange *et al.*, 2004) is a frequentist model selection approach for grouping algorithms, based on cross-validation. It has been demonstrated to perform competitively compared to a wide range of published cluster validation procedures (Lange *et al.*, 2004). The stability algorithm is a wrapper method for a clustering algorithm specified by the

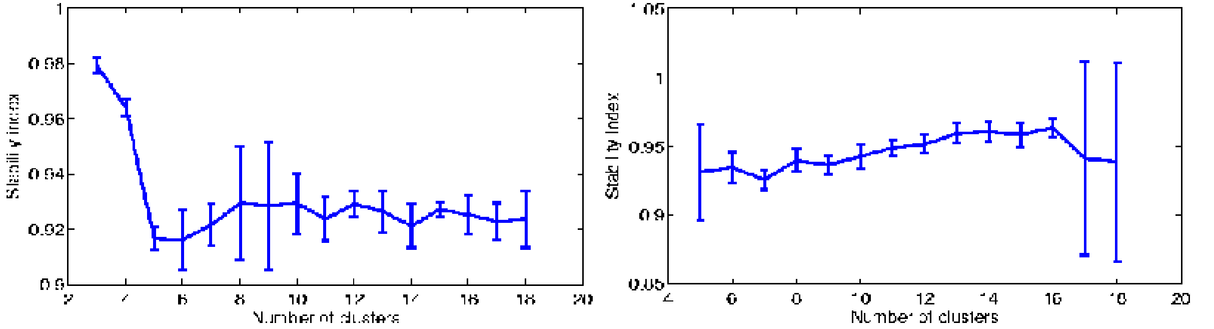


Figure 4.12: Stability index results over number of clusters, plotted for images in Fig. 4.8 (left) and Fig. 4.4 (right).

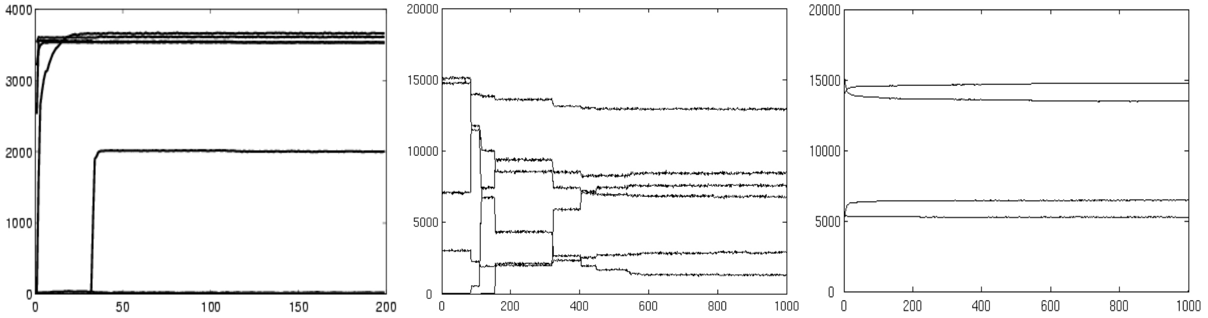


Figure 4.13: Cluster splitting behavior of the sampler for different images and parameters. The number n_k of sites assigned to each cluster (vertical) are drawn against the number of iterations (horizontal), with each graph representing a cluster. Left: Mondrian image (Fig. 4.1), no smoothing. Middle: Radar image (Fig. 4.4), no smoothing. Right: Radar image (Fig. 4.4), with smoothing.

user. It is applicable to any clustering algorithm which computes a unique assignment of an object to a cluster, e. g. it can be applied to a density estimate (such as mixture model algorithms) with maximum a posteriori assignments. The validation procedure works as follows: The set of input data is split into two subsets at random, and the clustering algorithm is run on both subsets. The model computed by the clustering algorithm on the first set (training data) is then used to *predict* a solution on the second set (test data). The two solutions on the second set, one obtained by clustering and one by prediction, are compared to compute a "stability index". The index measures how well the predicted solution matches the computed one; the mismatch probability is estimated by averaging over a series of random split experiments. Finally, the number of clusters is selected by choosing the solution most stable according to the index.

The DPM model is built around a Bayesian mixture model, consisting of the multinomial likelihood F and the Dirichlet prior distribution G_0. The Bayesian mixture without the DP prior can be used as a clustering model for a fixed number of segments. Inference of this model may be conducted by a MCMC sampling algorithm closely related to MacEachern's algorithm for DPM inference. The only substantial difference between the algorithms is the additional assignment probability term corresponding to the base measure, as observed in (Robert, 1995). A wrapper method like stability allows us to compare the behavior of the DPM approach to a method using exactly the same parametric model, including the base measure and its scatter parameter β. Only the parameter α is removed from the overall model, and the random sampling of the model order replaced by a search over different numbers of clusters.

Stability index results are shown in Fig. 4.12 for two images, the monkey image in Fig. 4.8 and the SAR image in Fig. 4.4. Results are not smoothed, because the subsampling strategy will break neighborhoods. In both cases, model order selection results for these noisy images are ambiguous. For the monkey image (upper graph), results for $K \geq 5$ are mostly within error bars of each other. A smaller number of clusters is ruled out, which is consistent with the unsmoothed DPM results (Tab. 4.1). For the SAR image, stability results are also ambiguous, but exhibit a significant, monotonous growth with the number of clusters, which is consistent with the monotonous behavior or the DPM results as α increases.

In general, stability has been reported to produce somewhat conservative estimates, since only the stability index of a solution is taken into account (Lange *et al.*, 2004). This observation is apparently reflected by the behavior of both methods on the monkey image, where the DPM approach settles at 6 clusters (with very small standard deviation), whereas stability advocates solutions with $K \geq 5$.

Convergence behavior

Gibbs sampling algorithms are notoriously slow, and it is often difficult to determine whether or not the algorithm has converged to the distribution of interest. Gibbs sampling results reported in the DPM literature are typically based on several thousand iterations.

To the advantage of our algorithm, we are interested in segmentation results rather than parameter estimates. The cluster labels are discrete and tend to stabilize after the initial burn-in. Therefore, after discarding the burn-in, class assignments can be estimated reliably from a small number of samples. The indicator for convergence used in the experiments is the relative fluctuation of class labels per iteration. The burn-in phase is assumed to be over once the number of assignments changed per iteration remains stable below 1% of the total number of sites. For the non-smoothing DPM sampler, this condition is usually met after no more than 500-1000 iterations – details depending on the input data and the scatter of the DP. These figures are comparable to those reported in the DPM literature. For example, (MacEachern, 1994) discards 1000 iterations as burn-in (and estimates are then obtained from 30000 subsequent iterations).

Fig. 4.13 shows the behavior of class assignment during the sampling process, for the noisy Mondrian and one radar image. For the Mondrian image with its well-separated segments, 40 iterations suffice for the clustering solution to stabilize (the cluster graph turns constant). On the radar image, both the non-smoothing and the smoothing version of the algorithm take about 600 iterations to stabilize, but their splitting behavior differs significantly: The standard DPM algorithm creates the last new significant cluster after about 150 iterations, while the DPM/MRF algorithm creates its classes during the first few iterations and slowly adjusts assignments throughout the sampling process. Without smoothing, large batches of sites are suddenly reassigned from one cluster to another (visible as jumps in the diagram). With smoothness constraints, clusters change gradually. Since the curves represent cluster sizes, they do not indicate the explorative behavior of the sampler. Even if the curve is smooth, the sampler may still explore a large number of possible states in parameter space, depending on the posterior.

4.1.7 Discussion

Segmentation models for mid-level vision have to address the two core issues of what a suitable model for individual segments should capture and how many segments should be inferred from an image. The last decade has seen significant progress in segmentation algorithms ranging from graph-based methods like partitioning models (Geman *et al.*, 1990), pairwise clustering (Hofmann and Buhmann, 1997) and Normalized Cut (Shi and Malik, 2000) to variational (Morel and Solimini, 1995) and statistical (Tu and Zhu, 2002) approaches. The specific nature of the images and the intended computer vision task most often determine the appropriateness of a model and the success of its related algorithm. The comparison is still subjective to a large degree, although the Berkeley data base of hand segmented natural color images (Martin *et al.*, 2004) allows us to benchmark new algorithms against human performance.

The applicability of the spatially constrained model is not restricted to either image segmentation or histogram clustering. Any kind of parametric mixture model may be used, by choosing the likelihood function F appropriately, and defining a suitable base measure to generate the parameter values. One might, for example, consider a k-means model with variable number of clusters and smoothness constraints, by defining F to be a Gaussian of fixed scale. The mean parameters are drawn from the base measure. If the base measure is also defined as a Gaussian (and therefore conjugate to F), the sampling algorithm proposed in Sec. 4.1.3 remains applicable as well. We expect that our model covers a large part of the landscape of segmentation algorithms since normalized cut and pairwise clustering can be written as weighted and unweighted versions of k-means in feature space (Roth *et al.*, 2003).

DPM methods do not “solve” the model order selection problem, because the number of clusters is replaced rather than removed as an input parameter. The utility of DP priors is not a decrease in the number of parameters, but the substitution of the constant model order by a random variable. The behavior of the random variable is parameter-controlled, and its eventual value estimated from data. Rather than specifying a number of image segments, the user can specify a *level of resolution* for the resulting segmentation. Part of the appeal of DPM-based models is their simplicity. Despite lengthy theoretical derivations, the final form of the model relevant for application is essentially a parametric mixture model with an additional term defined by the base measure. Familiar intuition for parametric mixture models remains applicable, and inference can be conducted by a sampling algorithm with a structure reminiscent of the expectation-maximization algorithm.

Since DPM and DPM/MRF models are built around a parametric model, careful parametric modeling is crucial for their successful application. The DP parameter α specifies a sensitivity with which the DP prior reacts to disparities in the data by creating additional clusters. The disparities are measured by the parametric model. As discussed in Sec. 4.1.6, modification of the parametric model will directly influence the DPM results. Hence, a DPM model can only be expected to work well for clustering if the class structure in the features is properly resolved by the parametric model. A clearly discernible cluster structure results in stable model order selection. Smoothing constraints can serve to emphasize cluster structure and stabilize results.

4.2 Model Order Adaptation

The previous section considered segmentation of individual images. Arguably, DPM-based models will develop their full potential when applied to multiple instances, for example, collections of radar images all generated by the same satellite. The problem considered in the following are video sequences, which consist of multiple similar instances, but with additional structure to be accounted for by the model. The number of segments may vary from image to image, but the images are drawn from the same source or very similar sources. If the number of segments is an input parameter, it has to be reset manually for each instance. Bayesian DPM models treat the number of segments as a random variable, with a distribution depending on the image instance. Since the distribution is controlled by parameters, they enable the data analyst to specify a segment resolution, possibly by calibrating the model parameters on a small subset of the data. Applied to new image instances with similar statistical properties, the model will automatically adapt the number of segments to variations in the data.

Application to large numbers of image instances requires efficient inference algorithms. An efficient Gibbs sampler similar to the one for DP mixture models is developed for the the time series model. To facilitate application of our model to the large amounts of data arising in video segmentation, we (i) show how the efficiency of the Gibbs sampler can be substantially increased by exploiting temporal smoothness and (ii) introduce a multiscale sampling method to speed up processing of individual frames. Just as the model, the multiscale algorithm is based on the properties of sufficient statistics.

4.2.1 Video Segmentation Problems

When clustering is applied to perform segmentation, the input data is typically a digital image (group the image into spatially coherent segments) or a time series (decompose the series into coherent segments along the time axis, such as speaker clustering). A different problem arises when video segmentation is formalized as a clustering problem: Given is a time series of fixed-size data frames, each of which has a spatial structure, i. e. the 2D structure of the frame image. The series is to be decomposed into a sequence of spatially coherent segmentations of the frames. The segmentation solution should reflect the temporal smoothness of the sequence. Clustering problems of this type have been actively studied in the video segmentation literature (see Tekalp, 2000, for an overview). For example, Weiss and Adelson (1996) propose a parametric mixture model for optical flow features with neighborhood constraints. The number of clusters is selected by a likelihood heuristic. Temporal context is modeled implicitly by using differential motion features. Explicit context models include designs based on HMMs (Bregler, 1997) or frame-to-frame model adaptation (Khan and Shah, 2001). A method which approaches the problem's time series structure in a manner similar to Bayesian forecasting has recently been suggested by Goldberger and Greenspan (2006). The authors propose a Gaussian mixture model to represent image rather than motion features. Temporal context is incorporated by using the estimate obtained on a given frame in the sequence as prior information for the following frame.

4.2.2 Setting and Notation

Much as in the previous section, we assume data $\mathbf{x}^t := (\mathbf{x}_1^t, \dots, \mathbf{x}_n^t)$ to be measured at the sites of an image (with $i = 1, \dots, n$ indexing the sites). A video sequence consists of a set of images, indexed along the time axis by $t = 1, 2, \dots$. Site locations within the image are assumed to remain constant for all t. Once again, we assume that each frame image $\mathbf{x}^t$ decomposes into a number K^t of segments, which may change from frame to frame. A clustering solution for each frame is encoded by assignment variables $\mathbf{Z}^t := (Z_1^t, \dots, Z_n^t)$, where $Z_i^t \in \{1, \dots, K^t\}$. That is, the number n of assignment variables per frame remains constant over time, but the range may vary. At each time step t, segmentation the respective image $\mathbf{x}^t$ is modeled by a mixture with parametric component distributions $F(\mathbf{x}_i^t|\theta_i^t)$. Notation therefore remains largely identical to the previous section, except for the additional time index.

In image and video processing applications, the input data usually consists of multiple channels. For standard color videos, three channels correspond to the three color space dimensions. Additional channels may include other features in the form of transformed data or filter responses. For multiple data channels indexed $c = 1, \dots, C$, multiple observations $(\mathbf{x}_{i,c}^t)_{c=1,\dots,C}$ are obtained for each frame t and location i. These are represented in the model as a product of likelihoods. That is, the generative model is obtained by substituting suitable product distributions $\prod_{c=1}^{C} F(\mathbf{x}_{i,c}^{t+1}|\theta_{i,c}^{t+1})$ and $\prod_{c=1}^{C} G_c^{t+1}$ for F and G_0 in (4.2.12).

4.2.3 Clustering Model

Clustering solutions for the time series problem are required to temporally coherent. In a video sequence, each time step corresponds to a single frame image. The overall clustering solution then consists of a segmentation for each frame. If the number of clusters can change between frames, a suitable clustering method must be *order-adaptive*, i. e. capable of adjusting the model order K^t over time. Order-adaptive methods require (i) automatic model order selection and (ii) a meaningful way to match clusters between frames. If clustering solutions are obtained independently on each frame, the cluster correspondence problem must be addressed by matching heuristics. Any principled approach requires the use of context information, i. e. the clustering solution for a given frame has to be obtained in a manner conditional on the solutions for the previous frame. In this section, we discuss how cluster structure can be propagated along a time series if the clustering solutions on individual frames are controlled by a DP prior.

For temporal coherence, we require that, if $Z_i^t = k$, then also $Z_i^{t+1} = k$ with high probability, unless the corresponding observations $\mathbf{x}_i^t$ and $\mathbf{x}_i^{t+1}$ differ significantly. For the video segmentation problem, this reflects the assumption that size and location of segments change slowly on the time scale of frame renewal. The standard Bayesian approach to address temporal coherence requirements in time series models is to encode context by priors. The posterior distribution of the model parameter vector θ^t at a given time is used as prior distribution for θ^{t+1}. This requires a conjugate model, assuming, the class of the prior distribution should not change between time steps.

Conjugate Models for Time Series

Conjugate models admit the use of a posterior under previous observations as a prior for future observations, in a manner similar to the interpretation of conjugate priors outlined in Sec. 2.3.2. Most time series data is not exchangeable, though, and conjugate models should not aggregate information over time in the manner of an exchangeable model.

Exchangeable data is the most straightforward case. Let $F(\,.\,|\theta)$ be an exponential family observation model with sufficient statistic s, and $G(\,.\,|\lambda, \mathbf{y})$ its conjugate prior. If observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are generated i.i.d.

from $F(\,.\,|\theta)$, they are conditionally independent given θ. For θ unknown, the $\mathbf{x}_i$ are not independent, but exchangeable by conditional independence. In this case, the posterior under prior $G(\,.\,|\lambda, \mathbf{y})$ is

$$G\Big(\theta\Big|\lambda + n, \mathbf{y} + \sum_i s(\mathbf{x}_i)\Big) \propto \Big(\prod_{i=1}^{n} F(\mathbf{x}_i|\theta)\Big) G(\,.\,|\lambda, \mathbf{y}) \,. \tag{4.2.1}$$

The posterior mean is $\mathbb{E}_{\Theta|x_1,\ldots,x_n}[\theta] = \frac{\mathbf{y}+\sum s(\mathbf{x}_i)}{\lambda+n}$ i. e. the sample mean of $s(X_i)$ up to an offset $\mathbf{y}$. The scalar $\lambda + n$ controls the posterior scatter, and as n grows large, the posterior will peak sharply at its mean, which for large n stabilizes around $\mathbb{E}[s(X)]$. Once the mean has stabilized, additional observations will affect the posterior mainly by sharpening the peak through increase of n. The conjugacy implies a natural chaining property. If, for example, the observations are split into two subsets $\mathbf{x}_1, \ldots, \mathbf{x}_{n_0}$ and $\mathbf{x}_{n_0+1}, \ldots, \mathbf{x}_n$, the posterior can be equivalently computed at once or in two steps:

$$G\Big(\theta\Big|\lambda + n, \mathbf{y} + \sum_{i=1}^{n} s(\mathbf{x}_i)\Big) = \frac{\Big(\prod_{i=n_0+1}^{n} F(\mathbf{x}_i|\theta)\Big)}{F(\mathbf{x}_{n_0+1}, \ldots, \mathbf{x}_n)} G\Big(\theta\Big|\lambda + n, \mathbf{y} + \sum_{i=1}^{n_0} s(\mathbf{x}_i)\Big) \,. \tag{4.2.2}$$

More generally, observations can be assumed to be acquired one by one, with the posterior updated on each measurement. The resulting final posterior will be identical to the posterior obtained from the whole set of observations at once.

To account for data that is not exchangeable along the time axis, such conjugate chaining is still applicable, if the data is not aggregated indefinitely. For the video segmentation problem, a suitable modeling assumption is that changes between consecutive time steps are small, and that the current state (the current image) is the best available guess of what the next image may look like. By the Diaconis-Ylvisaker characterization of conjugacy (Theorem 19), for prior $G(\,.\,|\lambda, \mathbf{y})$ and a single observation $\mathbf{x}^t$, the mean of the posterior[1] is:

$$\mathbb{E}_{\mu_{\Theta|X^t=\mathbf{x}^t}}\Big[\mathbb{E}_{s(X^{t+1})|\theta}\big[s(X^{t+1})|\theta\big]\Big] = \frac{y + s(x^t)}{\lambda + 1} \,. \tag{4.2.3}$$

If $\mathbf{y}$ is set to the origin, $s(x^t)$ is reproduced in expectation in the limit $\lambda \searrow 0$. Otherwise, the observation $s(x^t)$ is offset by $\mathbf{y}$. The offset is not necessarily

[1] Eq. 4.2.3 holds only if the inner expectation on the left-hand side is taken w. r. t. the random variable $s(X^{t+1})$. If the expectation is computed w. r. t. X^{t+1}, the relation becomes nonlinear.

interpretable as a translation. For estimation of a multinomial parameter vector under a Dirichlet distribution prior, for example, $\mathbf{y}$ represents a finite probability distribution. The image $s(\mathbf{x}^t)$ of an observation $\mathbf{x}^t$ is a finite probability distribution as well, possibly up to normalization. The right-hand side of (4.2.3) interpolates between $\mathbf{y}$ and $s(x^t)$ in the probability simplex. If $\mathbf{y}$ is chosen as uniform (the center point of the simplex), it will act as a regularizer on $s(\mathbf{x}^t)$. Setting $\mathbf{y}$ to an extremal point of the simplex will emphasize the respective event in θ, etc (cf also the discussion in Sec. 4.1.6 for the Dirichlet case). In general, $\mathbf{y}$ can be interpreted as a prior guess for θ, against which observations are interpolated.

Eq. (4.2.3) implies that a model using the posterior

$$G(\theta^t|\lambda+1, \mathbf{y}+s(\mathbf{x}^t)) \propto F(\mathbf{x}^t|\theta^t)G(\theta^t|\lambda,\mathbf{y}) \tag{4.2.4}$$

as prior for θ^{t+1} will indeed rely on the current state $s(\mathbf{x}^t)$ as best available guess, and smooth by interpolating against $\mathbf{y}$. The model sequentially generates data according to

$$\begin{tikzcd}
F(\mathbf{x}_i^1|\theta_i^1) \arrow[rd] & F(\mathbf{x}_i^2|\theta_i^2) \arrow[rd] & F(\mathbf{x}_i^3|\theta_i^3) & \cdots \\
G(\theta^1|\lambda,\mathbf{y}) \arrow[u] \arrow[r] & G(\theta^2|\lambda+1,\mathbf{y}+s(\mathbf{x}^1)) \arrow[u] \arrow[r] & G(\theta^3|\lambda+1,\mathbf{y}+s(\mathbf{x}^2)) \arrow[u] &
\end{tikzcd}$$

(4.2.5)

If the data is approximated reasonably well by an exchangeable model over a small time window, one may variate upon the chaining strategy by accumulating a fixed number of observations, as

$$G\Big(\theta^{t+1}\Big|\lambda+\tau, \mathbf{y}+\sum_{t-\tau<r\leq t} s(\mathbf{x}_r)\Big) \propto \prod_{t-\tau<r\leq t} F(\mathbf{x}^r|\theta^r)G(\theta^{t-\tau+1}|\lambda,\mathbf{y}) \tag{4.2.6}$$

The window equips the model process with a τ-step memory.

Remark 28 (Time series interpretation of the model)**.** Since $\mathbb{E}\left[\Theta^{t+1}\right] = \frac{\mathbf{y}+s(x^t)}{\lambda+1} = c\mathbf{y} + c\cdot s(x^t)$, the value assumed by Θ^{t+1} is representable as

$$\theta^{t+1} = c\mathbf{y} + c\cdot s(x^t) + \varepsilon^t \ , \tag{4.2.7}$$

where $\varepsilon^t \sim G(\,.\,|\lambda+1, 0)$. If $s(\mathbf{x}^t)$ was replaced by θ^t, the model would be AR(1). If θ represents a Gaussian mean (under a normal prior), ϵ^t is zero-mean white noise, and the model is covariance-stationary, because

$c < 1$. In general, note that this time series lives in parameter space. It has a systematic drift for $\mathbf{y} \neq 0$. As discussed above, the drift may be regarded as e. g. a bias towards a regular solution, depending on the choice of $\mathbf{y}$. The actual model substitutes an observation $s(x^t)$ conditional on θ^t for the value of θ^t. It is worth noting that such linear processes in parameter space capture all information relevant for the model: In conjugate models, whatever happens in parameter space is linear, and any non-linearities in the data are expressed by means of the map s.

Conjugate Chaining of Dirichlet Process Priors

In order to propagate cluster structure (in terms of a Bayesian mixture model) along the time axis, the model has to generate mixing distributions μ_Z as in (2.5.1). That is, the parameter Θ^t in the model above takes values in the set of mixing distributions:

$$\begin{aligned} \mathbf{x} \sim p(\mathbf{x}) &= \int F(\mathbf{x}|z) d\mu_{\mathrm{Z}}(z) \\ \mu_{\mathrm{Z}} \sim G & \end{aligned} \tag{4.2.8}$$

Once again, we will write θ and θ_k^* for the values assumed by Z, such that $\theta \sim \mu_{\mathrm{Z}}$ and the mixture components F are parameterized by θ or θ_k^* (cf. Rem. 27 on notation). For finite Bayesian mixtures, G is a suitable parametric product prior (cf. Sec. 2.5.2), and μ_{Z} is generated in form of a density m_Z as in (2.5.4), by generating the parameters θ_k^* and c_k. In a nonparametric setting, μ_{Z} is drawn from a Dirichlet process substituted for G. The Dirichlet process has a natural conjugate property. It will be discussed in detail in Chap. 5, but is inherent in Ferguson's original characterization of the DP posterior:

$$\theta_1, \dots, \theta_n \sim \mathrm{DP}(\alpha G_0) \qquad \Longrightarrow \qquad \theta_{n+1} \sim \mathrm{DP}\Big(\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}\Big) . \tag{4.2.9}$$

The conjugate chaining of priors and posteriors, as discussed above, therefore carries over immediately to the nonparametric case (as does the concept of sufficiency and any intuition based upon it, cf. Chap. 5).

If $\theta_1, \dots, \theta_n$ is a sample of size n from a DP, i. e. if $m_{\mathrm{Z}} \sim \mathrm{DP}$ and $\theta_i \sim m_{\mathrm{Z}}$, we will write

$$\hat{m}_n := \frac{1}{n} \sum_{i+1}^{n} \delta_{\theta_i} . \tag{4.2.10}$$

The nonparametric analogue of prior-posterior chain defined in (4.2.4) for the parametric case is then

$$m_{\mathrm{z}}^{t+1} \sim \mathrm{DP}\left(\alpha G_0 + \hat{m}_n^t\right) \ . \tag{4.2.11}$$

That is, α is substituted for λ, G_0 for y and $\hat{m}_n^t$ for $s(\mathbf{x}^t)$. The overall data generation may be summarized as

$$\begin{aligned} \mathbf{x}_i^{t+1} &\sim F(\,.\,|\theta_i^{t+1}) \qquad\qquad\qquad\qquad\qquad\qquad (4.2.12) \\ \theta_1^{t+1},\ldots,\theta_n^{t+1} &\sim G^{t+1} \qquad \text{and} \qquad \hat{m}_n^{t+1} := \frac{1}{n}\sum_{i+1}^{n} \delta_{\theta_i^{t+1}} \\ G^{t+1} &\sim \mathrm{DP}\left(\alpha G_0 + G^t\right) \ . \end{aligned}$$

In the diagram representation (4.2.5), an additional layer is added for the mixture:

$$\begin{tikzcd}
F(\mathbf{x}_i^1|\theta_i^1) & F(\mathbf{x}_i^2|\theta_i^2) & F(\mathbf{x}_i^3|\theta_i^3) & \\
\theta_i^1 \sim m_{\mathrm{z}}^1 \arrow[u] \arrow[dr] & \theta_i^2 \sim m_{\mathrm{z}}^2 \arrow[u] \arrow[dr] & \theta_i^3 \sim m_{\mathrm{z}}^3 \arrow[u] & \\
\mathrm{DP}\left(\alpha G_0\right) \arrow[u] \arrow[r] & \mathrm{DP}\left(\alpha G_0 + \hat{m}_n^1\right) \arrow[u] \arrow[r] & \mathrm{DP}\left(\alpha G_0 + \hat{m}_n^2\right) \arrow[u] \arrow[r] & \cdots
\end{tikzcd} \tag{4.2.13}$$

Once again, the model implies exchangeability of observations at each time step, but not along the sequence. If, for t fixed, multiple instances of m_{z}^t are generated, these will be exchangeable. However, actual data is assumed to be generated by a single draw from each time step, and two such draws $m_{\mathrm{z}}^{t_1}$, $m_{\mathrm{z}}^{t_2}$ are not generally exchangeable unless $t_1 \neq t_2$.

4.2.4 MCMC Inference

Two questions will be addressed in the following: How the standard blocked Gibbs sampler for DPM inference can be adapted to perform inference along the time series model described above, and how efficiency of the sampler can be improved to cope with the large amount of data typically encountered in video segmentation or comparable problems, by exploiting the temporal and spatial coherence in the data.

Existing hidden-variable methods can be modified for time series inference by initializing inference for a given time step by the model state estimated for the previous step. Each cluster is then indexed uniquely throughout the time series. A good estimate obtained for a given frame will provide an almost-perfect initialization for the subsequent frame. If changes between frames are small, the task of the sampler is thereby reduced from approximating to tracking the evolving structure. To increase sampler efficiency for individual time steps, temporal tracking is combined with a multiscale algorithm.

Sampling in Time Series

Parameter inference for the time series clustering problem estimates the cluster parameters θ_k^t and the states of the assignment variables Z_i^t, for each $t = 1, \dots$. Estimates are obtained by sampling the relevant posteriors with a Gibbs sampler. To derive a suitable algorithm, we note that, for a given time index t, recovering the states Z_i^t and parameters θ_k^t given the current observations $\mathbf{x}_i^t$ is a DPM mixture inference problem with prior $\mathrm{DP}(\alpha G_0 + \hat{G}_n^t)$. The history of the process enters via the prior parameter, i. e. the measure $(\alpha G_0 + \hat{G}_n^t)$. Full conditionals for the Gibbs sampler are immediately obtained from the standard sampling algorithm, by substituting $(\alpha G_0 + \hat{G}_n^t)$ for αG_0. (A sampler for a series with a τ-step memory can be obtained by substituting the corresponding posterior parameter in (4.2.6)). Estimates for the whole time series can be computed by running the Gibbs sampler for the appropriate posterior at each time step. The parameter estimates (summarized by $\hat{G}_n^t$) are then substituted into the DP prior of the subsequent step. The algorithm is an online method, as it only performs a single pass over the time series. The cluster correspondence problem is solved implicitly, by propagating information from one time step to the next through the DP base measure. Initially, the same clusters as in the previous step are available for assignment, and their indices are preserved. Classes may be newly generated by drawing from the continuous component G_0 of the DPM, or deleted if no longer supported by the data. Gibbs sampling is potentially time-consuming, and performing a full run of a DPM Gibbs sampler for each time step in the series is computationally prohibitive. A substantial speed-up is achieved by exploiting temporal smoothness. If changes in the data occur slowly w. r. t. to the time scale (frame rate) of the time series, the model state estimated at time t provides an almost-perfect initialization for sampling at time $(t + 1)$. The algorithm therefore obtains an initial estimate at time $t = 1$ by performing a full run of the Gibbs sampler. For $t \geq 2$, the Gibbs sampler is initialized by the previous model state, and then run only for a few steps, to allow the model to adapt to changes in the data.

Multiscale Sampling

For data exhibiting a spatial neighborhood structure, DPM inference algorithms that are more efficient then the standard Gibbs sampler can be derived using a multiscale approach. Multiscale methods attempt to increase the efficiency of iterative algorithms by replacing the original input problem with a compressed replacement problem (*coarsening*). This reduced-size problem is solved, and the solution transformed into a solution of the larger input problem (*refinement*). The compression operation exploits neighborhood structures in the data (such as spatial or sequential neighborhoods). In images, adjacent pixels are grouped into blocks, and each block B is compressed by computing a summary variable $\mathbf{x}_B$. The coarsened problem is given by the set of summary variables for all blocks. The coarsening operation therefore has to be designed to limit the loss of relevant information under compression, and to result in a coarse-scale problem to which the processing algorithm in question is applicable.

Coarsening. Our aim is the design of MCMC sampling algorithms. The information to be preserved under coarsening is therefore the information relevant to statistical parameter estimation. A simple averaging approach is not suitable in general, as it will only preserve moment information of first order. For the models considered in our work, a suitable coarsening approach can be derived from the properties of sufficient statistics. For an exponential family density as in (2.2.15), all information relevant to parameter estimation is contained in the sufficient statistic $s(\mathbf{x})$. Furthermore, for multiple observations $\mathbf{x}_1, \dots, \mathbf{x}_n$, the sum $\hat{s}_n := \sum_{i=1}^{n} s(\mathbf{x}_i)$ is sufficient. Given a data block B, consisting of the observations $\{\mathbf{x}_{b_1}, \dots, \mathbf{x}_{b_N}\}$, the summary variable $\mathbf{s}_B$ is computed as

$$\mathbf{s}_B := \sum_{i=1}^{N} s(\mathbf{x}_{b_i}) \ . \tag{4.2.14}$$

If $\mathbb{R}^d$ data, for example, is modeled by a Gaussian distribution, the summary variable will be the pair $\mathbf{s}_B = \left(\sum_{i=1}^{N} \mathbf{x}_{b_i} \, , \, \sum_{i=1}^{N} \mathbf{x}_{b_i} \mathbf{x}_{b_i}^T\right)$. Coarsening is therefore performed by averaging in parameter space, in contrast to the standard multiscale schemes used by many computer vision algorithms, which average in the data domain. A spatial partition of the input data into blocks $B_1, \dots, B_m$ will result in a set of summary variables $\mathbf{x}_{B_1}, \dots, \mathbf{x}_{B_m}$. The DPM sampling algorithm described is directly applicable to this replacement data, by substituting summary variables $\mathbf{x}_B$ for sufficient statistic values $s(\mathbf{x}_i)$.

The coarsening operation is *perfect* in the sense that it does, by the properties of sufficient statistics, preserve all information relevant for estimation purposes. More precisely, assume that $\mathbf{x}_{b_1}, \ldots, \mathbf{x}_{b_N} \sim F(\,.\,|\theta)$, with a conjugate prior $G(\theta|\lambda, \mathbf{y})$ on the parameter. Then the posterior Π satisfies the invariance

$$\Pi(\theta|\mathbf{s}_B; \lambda, \mathbf{y}) = \Pi(\theta|\mathbf{x}_{b_1}, \ldots, \mathbf{x}_{b_N}; \lambda, \mathbf{y}) \,, \tag{4.2.15}$$

since

$$\Pi(\theta|\mathbf{x}_{b_1}, \ldots, \mathbf{x}_{b_N}; \lambda, \mathbf{y}) = G\Big(\theta \Big| \lambda\mathbf{y} + \sum_{i=1}^{N} s(\mathbf{x}_{b_i})\Big) = \Pi(\theta|\mathbf{s}_B; \lambda, \mathbf{y}) \,.$$

Intuitively, a parameter estimated from data is a valid description of the data on the fine scale or *any* coarsened scale.

Refinement. The coarsening strategy described above and subsequent sampling on the coarse scale will result in a DPM clustering solution defined by cluster parameters $\theta_1^*, \ldots, \theta_k^*$. Because these parameters also define an admissible clustering solution on the fine (original) scale of the problem, it is not necessary to explicitly propagate coarse-scale assignments to the fine scale. Instead, a solution of the fine-scale problem is obtained by substituting the estimates θ_k^* into the fine-scale model and performing a single assignment step. We note that, as another consequence of the parametric description applying simultaneously over different coarsening scales, the method is capable of incorporating locally adaptive coarsening/refinement strategies.

Coarse-scale artifacts. When applied to clustering, the multiscale sampler may erroneously produce additional classes at a coarse scale, if a block summarized by a single variable during coarsening overlaps the boundary between two segments. The resulting mixed distribution may distinctively differ from the average distribution of both segments, and thus produce an additional cluster. Such errors can be corrected by the fine-scale assignment step. To illustrate the behavior of the sampler, Fig. 4.14 shows estimation results obtained on a simple artificial image consisting of three block segments arranged in sequence (i. e. there are two boundaries between adjacent segments). Local windowed grayscale histograms are extracted as features. As clustering model, we apply a DPM model, with a multinomial likelihood F to account for the histograms. The cluster parameters θ_k^* can be interpreted as average histograms of the respective cluster. These average histograms are plotted for the true (generative) model on the left in Fig. 4.14. The multiscale algorithm is run on the data with a coarsening coefficient of 2, and the coarse-scale solution is compared to the artificial ground-truth. When sampling on a coarse scale, the algorithm models five classes, for which the class parameter vectors θ_k^* are plotted on the right. Clusters 2 and 4 are due to mixing of histograms at the segment boundaries.

When assignments are performed for the fine-scale histograms to the coarse-scale class parameters, all histograms are correctly assigned to their original three classes, and clusters 2 and 4 remain empty. That is, coarse-scale artifacts vanish during the fine-scale assignment. The algorithm benefits from the ability of the DP to create new clusters for the boundary points, without distorting the remaining cluster structure.

Multiscale approaches to Markov Chain sampling have been considered e. g. in (Higdon *et al.*, 2003). These methods are based on the idea that suitable coarse-scale formulations of a Markov chain may mix faster than the original chain, and use a coupled formulation integrating both chains. The aim is to reduce the number of iterations required for the algorithm to converge, while retaining accuracy. In contrast, our approach mainly aims at reducing the execution time of individual iterations. Keeping in mind the large amounts of data arising in visual processing and video, we trade in accuracy and statistical guarantees for speed. Though the coarsening operation is perfect for individual distributions, it will lose in accuracy when the coarsening blocks overlap segment boundaries, as shown in the example. In practice, we should not expect the fine-scale assignment step to correct all errors. The rationale for risking a loss of accuracy is that, for vision applications, we put more emphasis on speed and plausible results than on statistical guarantees.

4.2.5 Experiments

This section provides experimental results for the application of the model to video segmentation. Experiments were performed on both synthetic data and real-world data (sequences from the MPEG4 benchmark set).

Processing pipeline. Features are extracted from each frame image by placing an equidistant grid within the image. A local window is placed around each grid node i, pixel values are extracted from within the window, and collected in a histogram (denoted $\mathbf{x}_i^t$ in the previous sections). For color images, the method is applied individually to each color channel. The resulting set of features for each frame is a list of multiple histograms, indexed by their position i within the image. On this data, the inference algorithm described in Sec. 4.2.4 is applied. Inference is conducted single-pass, and is hence capable of online processing. For each time step t, the assignment variables Z_i^t describe the estimated segmentation (i. e. Z_i^t is interpreted as the segment index of site i). In the examples shown in Figs. 4.15-4.18, segment assignments have been color-coded.

Results. Synthetic data experiments were conducted to verify the method's capability to adjust the number of clusters. The artificial data

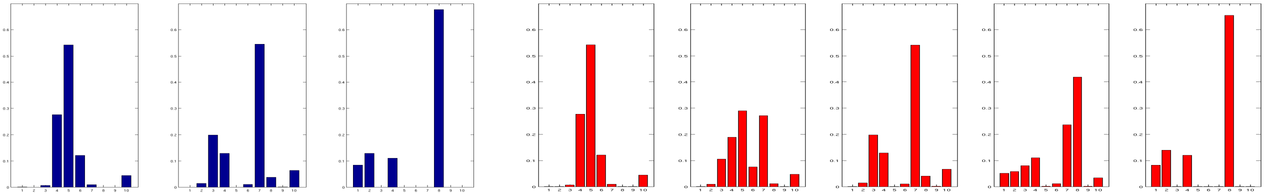


Figure 4.14: Erroneous creation of classes during coarse-scale sampling: Average histograms of the three input classes (blue/left) and parameter vectors of the five classes estimated by coarse-scale sampling (red/right).

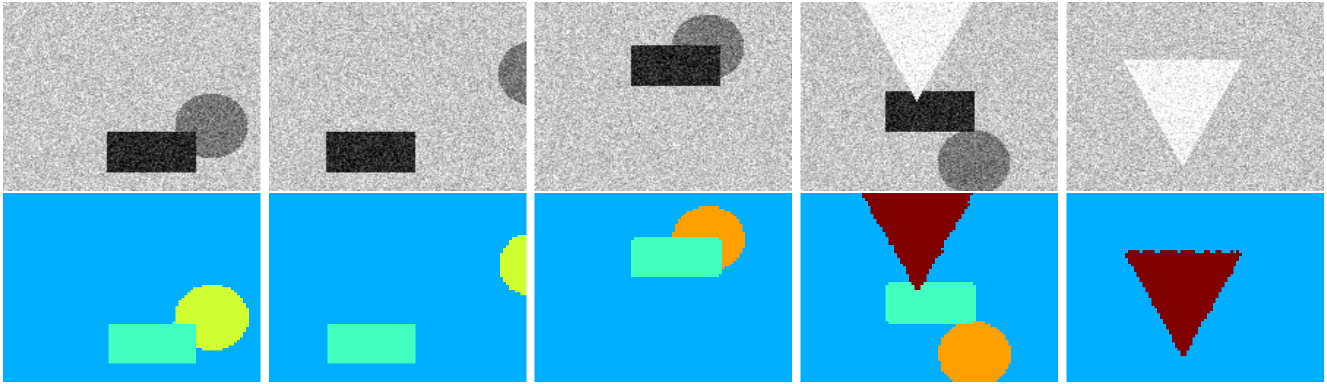


Figure 4.15: Synthetic noisy data (top) and segmentation results (bottom). Clusters are correctly created or deleted as objects enter or leave the scene.

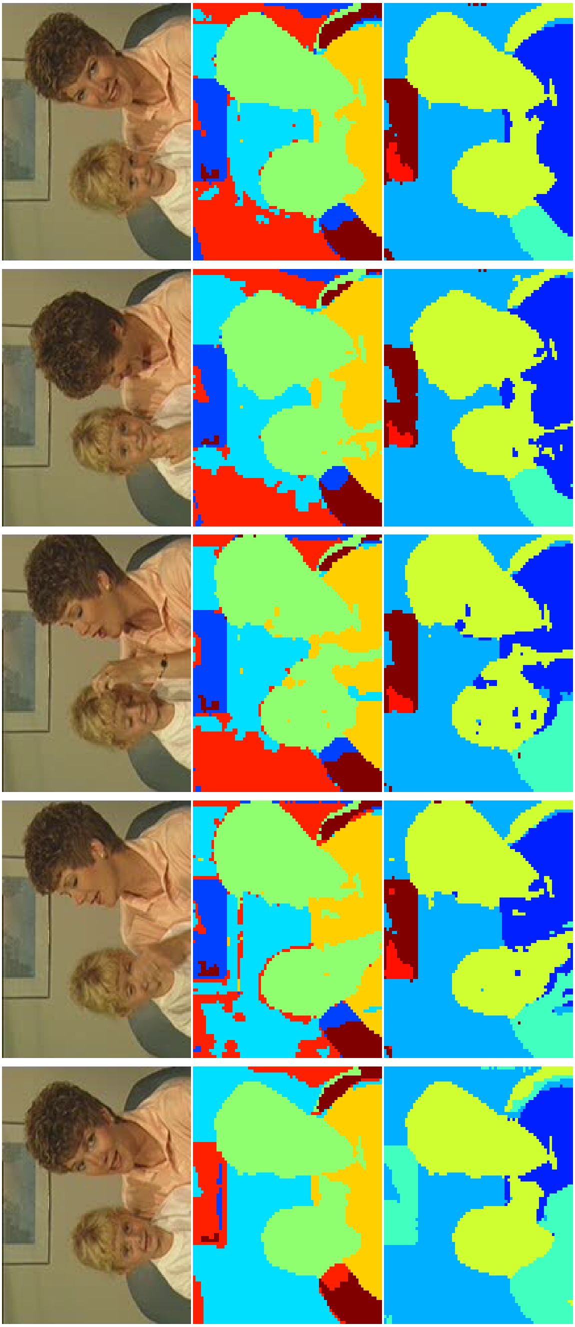


Figure 4.16: “Mother and child” test sequence (top). Results are shown for color and saturation histogram features (middle), and with additional location features modeled by Gaussian distributions (bottom).

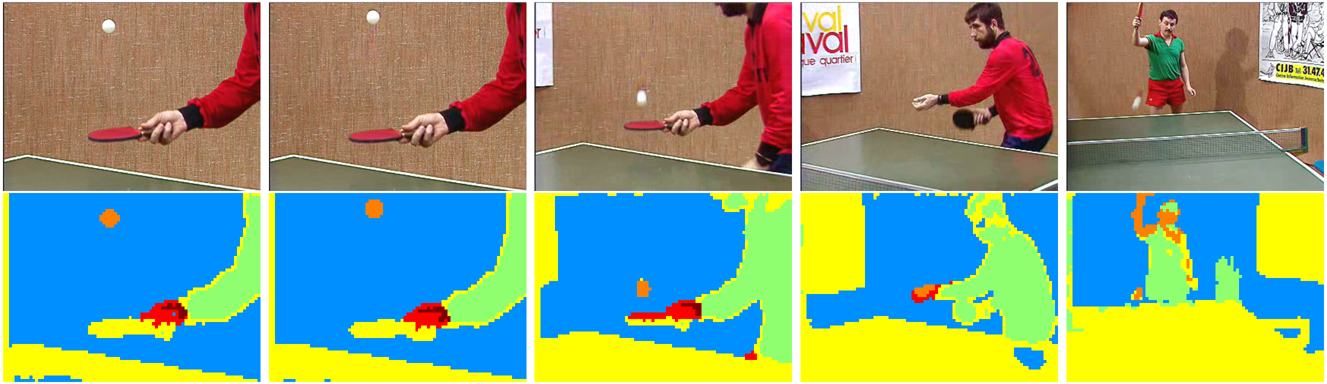


Figure 4.17: “Table tennis” test sequence, with histogram and location features.

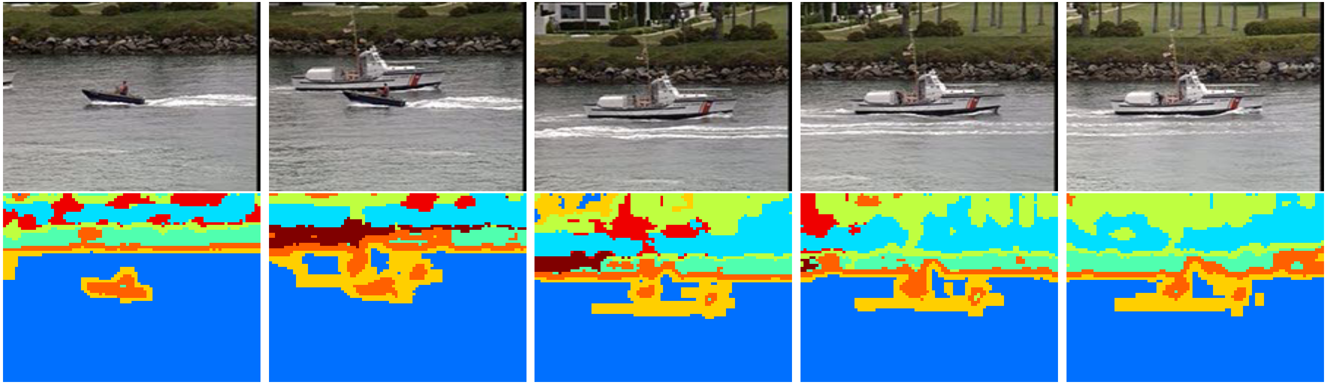


Figure 4.18: More difficult data:“Coastguard” test sequence (top) and segmentation results (bottom), with histogram and location features.

consists of simple geometric objects with additive Gaussian noise moving at random within a scene. Objects may newly appear or disappear, but only by entering or leaving the scene from the border (i. e. temporal changes are smooth). A sample experiment is shown in Fig. 4.15. Features used are local gray-scale histograms. In all experiments conducted, the algorithm consistently assigns each object to the same cluster over the whole running time of the sequence. The cluster number only changes if an object vanishes temporarily, either by occlusion or because it leaves the scene and reappears (as is the case for the disc in Fig. 4.15). As a mid-level vision algorithm, the method cannot (and should not) distinguish between initial appearance and reappearance of an object. Results on real video data were obtained on sequences from the MPEG4 benchmark set. Five feature channels where used: Four are histograms, representing the three RGB color channels plus saturation, and described by a multinomial likelihood F. In addition, we used a location feature, i. e. the center position of the local window in the image, represented by a two-dimensional Gaussian likelihood. The overall model likelihood is a product likelihood as described in Sec. 4.2.3. The scatter parameters of the parametric priors and the scatter parameter α of the DP, which control the level of cluster resolution, can be adjusted on the first few frames of the sequence. The key parameter of the feature extraction is the size of the local windows, which has to trade off sample size against precision: Large windows, to their advantage, contain many pixel examples, which results in stable histogram estimates and reduces scatter in feature space. Their drawback is a lack of precision: Large windows overlapping a cluster boundary generate histograms that represent a mixture of the two cluster distributions. Such mixtures tend to differ significantly from the individual distributions of the clusters, and hence cause additional clusters to appear at the segment boundaries. All results shown here were obtained using window sizes of 5×5 or 9×9 pixels. Sample results are shown in Figs. 4.16-4.18. In Fig. 4.16, results shown in the middle row where obtained using only the four histogram features (color and saturation). The background is split up into incoherent segments. Results are improved by the additional use of location features. Modeling these with a Gaussian in the spatial domain favors spatially coherent solutions, improving the segmentation of the background (bottom row). Likewise, all five features where used in the computation of results shown in Fig. 4.17. A more difficult sequence is shown in Fig. 4.18. In this case, local segmentation features provide poor information. Color differences within some segments (e. g. the large boat) are more significant than those between segments. With the size of the windows chosen sufficiently large during feature extraction to obtain stable input histograms, a boundary cluster effect is observable (note the two boats being split into an internal and a boundary segment). Results may possibly be improved by including additional motion features (such as histograms of frame differences). In general, the choice of features proves crucial for the performance of the segmentation algorithm. The parametric components of the clustering model (e. g. Gaussian and multinomial) are location-scatter type models, which represent “clouds” in their respective feature spaces. Like most mixture models, the method relies on the feature extraction step to map the segments to groups in feature space that are sufficiently well-separated to be resolved by the probabilistic model.

Average running times for our experiments on different video test sequences (300 frames at resolution 144 × 176 each) were: ~ 190 seconds at full resolution, ~ 110 seconds using a multi-scale sampler with coarsening coefficient 2, and ~ 35 seconds with a coarsening coefficient of 4. This does not include the feature extraction, i. e. the extraction of the fine-scale input data from the image sequence. The running time of the algorithm scales, in addition to the obvious dependence on the amount of input data, with the number of segments. The averages above were measured for relatively small numbers of classes ($K \leq 10$). If a large number of clusters is required (i. e. an over-segmentation), longer running times will have to be expected.

4.2.6 Discussion

The DP approach models the number of clusters as a random variable. The model order as an input parameter is replaced by a control parameter that allows the user to adjust the approximate level of cluster resolution. For fixed, static data sets, the approach may not constitute a practical advantage, as it arguable replaces one input parameter by another. We face a different situation for dynamic data, such as videos, where the number of clusters may change over time and the model has to adapt. Adaptation *requires* either a random description of the model order, or a transition heuristic (such as BIC scoring, or reversible jump in a Bayesian framework). Bayesian methods in general, and DP approaches in particular, are often regarded as inapplicable to data-intensive problems due to their computational costs. In our view, the reported results convincingly demonstrate that algorithmic efficiency need not pose an obstacle to the practical application of DP models, if temporal and spatial structure in the data can be exploited.

The achieved processing times of the sampler are just one order of magnitude below real time for videos in half-PAL format. We have not provided convergence results for the sampling algorithm, since our analysis of the coarsening operation implies that, for individual component distributions, results for standard Gibbs samplers carry over to the multiscale approach. Any inaccuracies are due to coarsening blocks overlapping segment boundaries, a problem hard to capture by mathematical analysis. Furthermore, our estimation results are necessarily approximate, since the sampler is only run for a few steps on each frame image. Such an algorithm, for which the observed data changes (smoothly) in regular intervals raises some interesting questions. On the one hand, frame changes may pose a problem, if the model has not yet been sufficiently adapted to the current data. On the other hand, small data perturbations may help to avoid local minima. In-depth analysis of the algorithm is beyond the scope of the present study.

The results presented here were computed on low-level features, such as color and saturation histograms, which necessarily results in limited precision. Since the model applies to both Gaussian and multinomial feature distributions, it is directly applicable to a wide range of features. Tracking applications, for example, require robustness but no coherent partition of the image. Hence, interest point features could be extracted on each frame and grouped with our model using a Gaussian likelihood supported on the frame image. Since DP models can be constrained by Markov random fields (cf. Sec. 4.1) the model itself can be extended by spatial smoothing, as advocated for video segmentation in (Weiss and Adelson, 1996). Such an extension would probably come at the price increased computation time, as more iterations per frame would be required for the smoothing to take effect.

Chapter 5

Construction of Nonparametric Bayesian Models

Ferguson's approach to the construction of a nonparametric Bayesian model applies Kolmogorov's extension theorem to a projective family of finite-dimensional Dirichlet distributions. The resulting projective limit measure is the Dirichlet process. Once the process measure is defined, its usefulness as a Bayesian prior is studied. Motivated by an increasing interest in the construction of process priors in the machine learning literature, we suggest a slightly different approach: Define a system of finite-dimensional Bayesian equations, with desirable properties, such as conjugacy; then apply the extension theorem to the entire system, in a manner that preserves all properties of importance.

The results established in the following are an attempt to make this approach reasonably generic. Our results describe a weak representation of an infinite-dimensional Bayesian model, in analogy to the weak distribution of a process measure in probability theory. We discuss in how far useful properties of the finite-dimensional system carry over to the projective limit case, and provide a number of construction examples.

5.1 Motivation

As far as the definition of new probability models is concerned, the primary focus of the nonparametric Bayesian community has been the construction of measures on the set of probability measures. The machine learning community, on the other hand, has traditionally been interested in Gaussian processes, an interest that was reinvigorated by both the popularity of kernel methods and the characterization of Gaussian processes as limits of certain neural networks by Neal (1994). Until recently, Gaussian process priors were regarded in machine learning mostly as an individual approach, rather than as one example of a larger class of Bayesian nonparametric methods. This perception has changed notably with the increasing popularity of the Dirichlet process.

Machine learning interest in the Dirichlet is motivated by clustering problems, and after DP models became known in the field, attempts at generalization of the model started to appear almost immediately. These were either combinations of DP models constructed in analogy to finite mixtures (such as the hierarchical DP of Teh *et al.* (2004) and the models described in Ch. 4), or based on generalizations and relatives of the DP available in the statistics literature (such as the Pitman-Yor process or Kingman's coalescent, see e. g. Teh *et al.* (2008a) for an application). The exception was a paper by Griffiths and Ghahramani (2005), which proposed a distribution of process type on a different class of infinite-dimensional objects, in this case infinite binary matrices. A number of constructions have since appeared in the machine learning literature, typically based on arguments involving an analytic limit. Roughly speaking, this point of view regards the DP as arising from a d-dimensional Dirichlet density in the limit $d \to \infty$. The density of some finite-dimensional model is written in a form explicitly or implicitly parameterized by the dimension of the underlying space, and a limit with respect to the dimension variable is computed. In some sense, one may argue, the limit is useful as an intuitive interpretation of the Dirichlet process. Unfortunately, it is also fundamentally flawed as a construction argument. Two pitfalls of the "density limit" idea are:

1. The carrier measure: The d-dimensional Dirichlet and Gaussian densities, for example, are defined w. r. t. Lebesgue measure, which has no extension to infinite-dimensional space. Even on infinite-dimensional separable Hilbert space, arguably the most benign infinite-dimensional space at our disposal, a Lebesgue measure does not exist (Skorohod, 1974). An analytic dimension limit of the density formula will lack meaning for lack of a carrier.
2. The infinite-dimensional measure is not guaranteed to admit a density representation. For example, the Dirichlet on the real line has not density. The set of Dirichlet posteriors on the real line is not dominated (App. C), and hence has no common representation as a conditional density.

The rigorous approach to the construction of infinite-dimensional measures is the Kolmogorov extension theorem. In the following, we will consider the construction of infinite-dimensional Bayesian models by means of the extension theorem. In particular, we will actually be interested in constructions taking into account conjugacy and sufficient statistics.

Conjugacy is a pervasive theme in the nonparametric Bayesian literature. Even in the finite-dimensional case, obtaining suitable exact or approximative expressions for the posterior of a Bayesian model is generally difficult, and analytic approaches tend to become much more subtle in infinite dimensions. The only known general form of exactly tractable posteriors are conjugate posteriors, and not surprisingly, all models considered in the nonparametric Bayesian literature, including GP and DP models, tailfree priors, Pólya trees and neutral to the right processes are of conjugate type (Walker *et al.*, 1999; Ghosh and Ramamoorthi, 2002).

The purpose will therefore be to construct models that admit conjugate posteriors. Since conjugacy is a property of pairs of priors and posteriors rather than of individual measures, it seems reasonable to consider the extension of complete Bayesian equations, rather than of individual measures. As Bayesian equations are in turn inherently parametric (whatever they may be called), this leads to the extension of conditional probabilities. Moreover, as was discussed to some detail in Ch. 2, there is a direct connection between conjugate Bayesian systems, exponential family models and sufficient statistics. The approach we will follow is hence to consider Bayesian system on finite-dimensional spaces. We will ask (i) whether and how they may be extended to infinite dimensions and (ii) which implications conjugate and sufficiency properties of the finite-dimensional system have for the infinite-dimensional case.

5.2 Results: A Qualitative Overview

A first observation is that all marginal models of a conjugate nonparametric Bayesian model are conjugate. Though somewhat trivial, the implication for the construction of conjugate infinite-dimensional systems is that they can only be constructed from conjugate marginals. Where conjugate systems are concerned, this will essentially restrict our attention to exponential family marginals. The first basic consideration will be the projective limit extension of conditional probabilities. In particular, we consider the case where the finite-dimensional marginals are regular conditional distributions.

A family of regular conditional probabilities on all finite-dimensional subspaces of an infinite-dimensional space, satisfying suitable consistency properties, define a conditional probability on the infinite-dimensional space. If the dimension is countable, the infinite-dimensional conditional probability is regular. (Proposition 36, p. 167.)

The consistency condition mentioned here is in principle a projection condition analogous to that required by the standard extension theorem. However, a conditional distribution in general carries less information than the original measure, and the projection condition has to account for a suitable structure in both arguments. An additional, recurrent theme will be conditions that guarantee regularity of the extended conditional measure. A non-regular conditional probability is not guaranteed to constitute a probability measure for (almost) all values of the condition. Non-regular conditionals are not suitable for Bayesian problems, because the posterior is not guaranteed to be a probability distribution (almost surely).

Given a notion of a projective limit for (regular) conditional probabilities, the extension approach can be applied to Bayesian systems:

A projective family of Bayesian models on all finite-dimensional subspaces of a space of countably infinite dimension uniquely defines an infinite-dimensional Bayesian model on the overall space. (Corollary 38, p.169.)

Sufficient statistics in conjugate models have, at least, two important aspects: One is that they characterize the probabilistic model. The second is that they define the parameter update by observation which defines the posterior distribution, as given by Eq. (2.3.20) for the finite-dimensional case. If the model is conjugate and the sufficient statistic known, the posterior is known explicitly. The immediate question is how a sufficient statistic can be extended along with the probability model.

In countably infinite dimensions, the sufficient statistics of a projective family of finite-dimensional Bayesian models define a sufficient statistic for the projective limit model. (Proposition 40, p. 169.)

Given the discussion so far and the result sketched above, the fact that conjugacy carries over to the infinite-dimensional case when the other results are applicable is almost inevitable.

In countably infinite dimensions, the projective limit of a conjugate system of finite-dimensional Bayesian models is conjugate. (Corollary 39, p. 169.)

The countability condition on the dimension is imposed to guarantee this regularity. This requirement is, at least at first glance, quite a nuisance: If the random quantity to be generated can in any way be thought of as a function, the set of dimensions is the function's domain. For a Gaussian or Dirichlet process to generate functions or probability distributions on the real line, respectively, the set of dimensions of the extension space must be isomorphic to $\mathbb{R}$, hence uncountable. For the special case of exponential family marginals even uncountable constructions define regular conditionals.

A nonparametric Bayesian system admits a conjugate posterior if the sampling distribution (likelihood) of all its finite-dimensional marginals is an exponential family model. In this case, the infinite-dimensional limit system is regular, even if the dimension is not countable. (Remark 33, p. 164 and proposition 43, p. 172.)

5.3 Formal Framework

The extension from finite- to infinite-dimensional measures uses the Kolmogorov extension theorem (cf Sec. 2.4.3). To apply the extension theorem, we have to define (i) product spaces with suitable topologies, (ii) a projection operation (from higher- to lower-dimensional spaces), and (iii) its preimage operation, which maps a set A to a cylinder set with base A. Application of extension techniques to Bayesian models requires their application to complete parametric model families, rather than individual measures, and we therefore have to consider conditional probabilities. Along with the projective (and therefore nested) structure of the sample space, we will have to require a nested structure on the σ-algebras on which the respective parametric models condition.

5.3.1 Process Measures

Random events will be modeled by the abstract probability space

$$(\Lambda, \mathcal{A}, \mathbb{P}) \tag{5.3.1}$$

with Λ a point set, $\mathcal{A}$ a σ-algebra on Λ and $\mathbb{P}$ a probability measure. Any measures considered in the following will be considered images of $\mathbb{P}$ under some random variable. We will generally denote random variables representing samples by X, and those representing a parameter by Θ. A conditional probability model representing the sampling distribution (or likelihood) is then of the form $\mu(X|\Theta)$.

Spaces and σ-algebras. Studying an infinite-dimensional model and its finite-dimensional marginals will require distinction between random variables of different dimensions, which take values in product spaces. As in the setting of the Kolmogorov extension theorem, product spaces are constructed from a Polish space Ω, with Borel σ-algebra $\mathcal{B}(\Omega)$. The (usually infinite) set of dimensions is denoted E, and E^* is the set of its finite subsets. The overall, infinite-dimensional product space and the corresponding product Borel system will be denoted

$$\Omega^{\mathrm{E}} := \prod_{i \in E} \Omega \qquad \text{and} \qquad \mathcal{B}^{\mathrm{E}} := \bigotimes_{i \in E} \mathcal{B}(\Omega) \; . \tag{5.3.2}$$

The finite-dimensional subspaces of Ω^{E} with their respective Borel algebras will be denoted $(\Omega^{\mathrm{I}}, \mathcal{B}^{\mathrm{I}})$ for $I \in E^*$. The sets I may be thought of as sets of axes indices.

It will prove important to distinguish the two σ-algebras $\mathcal{B}^{\mathrm{E}}$ and $\mathcal{B}(\Omega^{\mathrm{E}})$. The former is the E-fold product of the Borel sets on the component space Ω. This is the domain of measures constructed by means of Kolmogorov's extension theorem. $\mathcal{B}(\Omega^{\mathrm{E}})$, on the other hand, is the system of all Borel sets of the product space Ω^{E}, i. e. the domain the construction *should* be able to capture. The two cases coincide whenever E is countable. For uncountable E, however, only $\mathcal{B}^{\mathrm{E}} \subset \mathcal{B}(\Omega^{\mathrm{E}})$ holds. In particular, if Ω contains more than a single element, the product algebra does not contain the singletons. This can pose a problem for the construction of measures for Bayesian estimation, since sample observations are singletons.

Random variables. In analogy to the spaces, random variables will be indexed by a set of axes. For all $I \subseteq E$, write

$$X^{\mathrm{I}} : (\Lambda, \mathcal{A}) \rightarrow (\Omega^{\mathrm{I}}_x, \mathcal{B}^{\mathrm{I}}_x) \qquad \text{and} \qquad \mu^{\mathrm{I}}_X := X^{\mathrm{I}}(\mathbb{P}) \; . \tag{5.3.3}$$

In the construction of a Gaussian process, for example, we may choose $\Omega = \mathbb{R}$ and $E = \mathbb{Z}$. Then each X^{I} represents a random vector with $d = |I|$ entries in d-dimensional Euclidean space. The corresponding measure μ^{I}_X is a d-dimensional Gaussian distribution, representing the marginal on the subspace Ω^{I} of the infinite-dimensional Gaussian process distribution μ^{E}.

Projections and preimages. As in 2.4.3, the projection operator from Ω^{J}_x to Ω^{I}_x (with $I \subset J$) will be denoted $\mathrm{P}_{\mathrm{J},\mathrm{I}}$. For points $x^{\mathrm{J}} \in \Omega^{\mathrm{J}}_x$, it is explained as restriction of the list $x^{\mathrm{J}} = (x_i)_{i \in J}$ to $\mathrm{P}_{\mathrm{J},\mathrm{I}} x^{\mathrm{J}} := (x_i)_{i \in I}$, and for sets $A^{\mathrm{J}} \subset \Omega^{\mathrm{J}}_x$ as the set of all projections of points in A^{J}. The preimage under projection is denoted $\mathrm{R}_{\mathrm{J},\mathrm{I}} A^{\mathrm{I}} = \{x^{\mathrm{J}} \in \Omega^{\mathrm{J}}_x | \mathrm{P}_{\mathrm{J},\mathrm{I}} x^{\mathrm{J}} \in A^{\mathrm{I}}\}$. The projection of a measure is defined by means of a push-forward, as $(\mathrm{P}_{\mathrm{J},\mathrm{I}} \mu^{\mathrm{J}}_X)(A^{\mathrm{I}}) := \mu^{\mathrm{J}}_X(\mathrm{R}_{\mathrm{J},\mathrm{I}} A^{\mathrm{I}})$.

Example 29 (Gaussian process)**.** To illustrate the meaning and relations of the different objects, consider a Gaussian process distribution. Again choose $\Omega_x = \mathbb{R}$ and $E = \mathbb{Z}$. The Gaussian process generates infinite-dimensional random vectors in the space $\Omega_x^{\mathrm{E}} = \mathbb{R}^{\mathbb{Z}}$, with one entry for each element of $\mathbb{Z}$ and range $\mathbb{R}$ for each entry. The Gaussian process is given by the measure μ_X^{E}. For any finite subset $I \in E^*$ and corresponding finite-dimensional subspace Ω_x^{I}, the marginal on that subspace is

$$\mu_X^{\mathrm{I}} := \mathrm{P}_{\mathrm{J,I}}\mu^{\mathrm{E}} \; . \tag{5.3.4}$$

The definition of the Gaussian process states that all these marginals must be Gaussian. Projection operators are transitive, that is $\mathrm{P}_{\mathrm{K,I}} = \mathrm{P}_{\mathrm{K,J}} \circ \mathrm{P}_{\mathrm{J,I}}$ for $I \subset J \subset K$. Therefore, the distributions so defined satisfy $\mathrm{P}_{\mathrm{J,I}}\mu_X^{\mathrm{J}} = \mu_X^{\mathrm{I}}$ for any $I \subset J$, and form a projective family in the sense of Def. 22.

The situation typically considered in the following is that the infinite-dimensional measure μ_X^{E} is not known or given initially. Instead, a system of finite-dimensional marginals μ_X^{I} is given, which form a projective family. The existence of the process measure μ_X^{E} is then guaranteed by Kolomogorov's theorem, and any computation involving the behavior of μ_X^{E} on a finite-dimensional subspace can be performed in terms of the corresponding marginal.

5.3.2 Parametric Families of Process Measures

For the parameter variable corresponding to X^{I}, we will write Θ^{I}, hence

$$\Theta^{\mathrm{I}} : (\Lambda, \mathcal{A}) \to (\Omega_\theta^{\mathrm{I}}, \mathcal{B}_\theta^{\mathrm{I}}) \qquad \text{and} \qquad \mu_\Theta^{\mathrm{I}} := \Theta^{\mathrm{I}}(\mathbb{P}) \; . \tag{5.3.5}$$

The parametric model of X^{I} with parameter Θ^{I} is the Θ^{I}-conditional distribution $\mu^{\mathrm{I}}(X^{\mathrm{I}}|\Theta^{\mathrm{I}})$ of X^{I} (cf Sec. 2.2). All definitions given here apply equally for $I = E$ and for $I \in E^*$. For any measurable set $A^{\mathrm{I}} \in \mathcal{B}_x^{\mathrm{I}}$,

$$\mu^{\mathrm{I}}(A^{\mathrm{I}}|\Theta^{\mathrm{I}})(\omega) := \mu_X^{\mathrm{I}}(A^{\mathrm{I}}|\sigma(\Theta^{\mathrm{I}}))(\omega) = \mathbb{E}\left[\mathbb{I}_{A^{\mathrm{I}}}|\sigma(\Theta^{\mathrm{I}})\right](\omega) \; , \tag{5.3.6}$$

where the abstract conditional expectation $\mathbb{E}\left[\mathbb{I}_{A^{\mathrm{I}}}|\sigma(\Theta^{\mathrm{I}})\right]$ is taken w. r. t. μ_X^{I}. Whenever the sample space of X^{I} is Borel, the conditional probability will be assumed to be regular (i. e. a Markov kernel).

Projections of parametric models. So far, all notions discussed are fairly standard. However, to consider projective families of parametric models, we have to specify projections of parametric models. Assume that $\mu^{\mathrm{E}}(A^{\mathrm{E}}|\Theta^{\mathrm{E}})$ is regular. Then for almost all $\omega \in \Lambda$, the function $\mu^{\mathrm{E}}(\,.\,|\Theta^{\mathrm{E}})(\omega)$ is a measure on $(\Omega_x^{\mathrm{E}}, \mathcal{B}_x^{\mathrm{E}})$, and the standard projection operator $\mathrm{P}_{\mathrm{E,I}}$ can be applied, yielding a (projective) family of measures

$$\tilde{\mu}^{\mathrm{I}}(\,.\,) := \mathrm{P}_{\mathrm{E,I}}\mu^{\mathrm{E}}(\,.\,|\Theta^{\mathrm{E}})(\omega) \; . \tag{5.3.7}$$

It is straightforward to show (see below) that, if $\mu^{\mathrm{I}}(\,.\,|\Theta^{\mathrm{E}})(\omega)$ is the regular conditional probability of $\mu^{\mathrm{I}}_X = \mathrm{P}_{\mathrm{E,I}}\mu^{\mathrm{E}}_X$ with respect to Θ^{E}, then $\tilde{\mu}^{\mathrm{I}}(\,.\,) = \mu^{\mathrm{I}}(\,.\,|\Theta^{\mathrm{E}})(\omega)$. In other words, whether we first project and then condition or vice versa amounts to the same, and the diagram shown on the right commutes.

However, in both cases, the distribution is conditional on Θ^{E}, characterizing a parametric model in which all finite-dimensional marginals are parameterized by the infinite-dimensional parameter Θ^{E}. This does not reflect the structure of the parametric families we are interested in: The random variable X^{I} should depend only on the value of a restriction Θ^{I} of Θ^{E}. That is, the projection operator (or a corresponding operation) should be applied to Θ^{E} as well, and the parametric family of X^{I} should be fully specified by conditioning on the restriction Θ^{I}. In the terminology of Sec. 2.2.2, this implies sufficiency of $\sigma(\Theta^{\mathrm{I}})$ for X^{I}. A notion of projection adapted to this notion of multidimensional parametric models is given by the following definition.

$$\begin{tikzcd}
\mu^{\mathrm{E}}(\,.\,|\Theta^{\mathrm{E}}) \arrow[d, "\mathrm{P}_{\mathrm{E,I}}"'] & & \mu^{\mathrm{E}}_X \arrow[ll, "{\mathbb{E}[\mathbb{I}_{\{\,.\,\}}|\sigma(\Theta^{\mathrm{E}})]}"'] \arrow[d, "\mathrm{P}_{\mathrm{E,I}}"] \\
\mu^{\mathrm{I}}(\,.\,|\Theta^{\mathrm{E}}) & & \mu^{\mathrm{I}}_X \arrow[ll, "{\mathbb{E}[\mathbb{I}_{\{\,.\,\}}|\sigma(\Theta^{\mathrm{E}})]}"]
\end{tikzcd}$$

Definition 30 (Projector on conditional probabilities)**.** Let $\{\Omega^{\mathrm{I}}|I \in E^*\}$ be a system of product spaces and $X^{\mathrm{I}} : (\Lambda, \mathcal{A}) \to (\Omega^{\mathrm{I}}, \mathcal{B}^{\mathrm{I}})$ random variables with image measures $\mu^{\mathrm{I}} = X^{\mathrm{I}}(\mathbb{P})$. Let $\{\mathcal{C}^{\mathrm{I}}|I \in E^*, \mathcal{C}^{\mathrm{I}} \subset \mathcal{A}\}$ be a system of σ-subalgebras, such that $\mathcal{C}^{\mathrm{I}} \subset \mathcal{C}^{\mathrm{J}}$ whenever $I \subset J$. Then the projector $\mathrm{P}^*_{\mathrm{J,I}}$ on the conditional probabilities $\mu^{\mathrm{I}}(\,.\,|\mathcal{C}^{\mathrm{I}})$ will be defined as

$$\mathrm{P}^*_{\mathrm{J,I}}\mu^{\mathrm{J}}(A^{\mathrm{I}}|\mathcal{C}^{\mathrm{J}}) := \mu^{\mathrm{J}}(\mathrm{R}_{\mathrm{J,I}}A^{\mathrm{I}}|\mathcal{C}^{\mathrm{I}})\ . \tag{5.3.8}$$

For the parametric family case, the σ-subalgebras $\mathcal{C}^{\mathrm{I}}$ in the definition can be read as $\mathcal{C}^{\mathrm{I}} = \sigma(\Theta^{\mathrm{I}})$. Since $\mathcal{C}^{\mathrm{I}} \subset \mathcal{C}^{\mathrm{J}}$ whenever the projector $\mathrm{P}^*_{\mathrm{J,I}}$ is well-defined, the commutative structure shown above carries over to $\mathrm{P}^*_{\mathrm{J,I}}$.

Remark 31 (Index and projector notation for parameter spaces)**.** An additional complication concerning the parameter variable is that, depending on the chosen distribution model, the parameter for a given X^{I} may have a different number of dimensions than the observation. In the Gaussian case, for example, the distribution is parameterized by a d-vector and a $d \times d$-matrix, such that the parameter space for X^{I} should be indexed by $I \cup I \times I$ (neglecting the positive definiteness constraint on the matrix). Such cases can be treated within the same framework: For conditioning, parameter random variables Θ^{I} will be substituted for by their respective generated σ-algebras $\sigma(\Theta^{\mathrm{I}})$, and the only requirement to be ensured is that these σ-algebras are nested, in the sense that $\sigma(\Theta^{\mathrm{I}}) \subset \sigma(\Theta^{\mathrm{J}})$ if $I \subset J$. The notation used here is heavy on indices as it is, and quickly turns awkward for parameter spaces with complicated index sets. If, for example, the parameter space corresponding to Ω^{I}_x is a product space $(\Omega_\theta)^{\mathrm{I}\cup\mathrm{I}\times\mathrm{I}}$, a projector would be of the form $\mathrm{P}_{\mathrm{J}\cup\mathrm{J}\times\mathrm{J},\mathrm{I}\cup\mathrm{I}\times\mathrm{I}}$. Since the product $(\Omega_\theta)^{\mathrm{I}\cup\mathrm{I}\times\mathrm{I}}$ is completely determined by I, and to maintain some semblance of readability, we will instead write $\Omega^{\mathrm{I}}_\theta$ for the space and $\mathrm{P}_{\mathrm{J},\mathrm{I}}$ for the projector. That is, we will use the index I symbolically for parameters in such cases, indicating that $\Omega^{\mathrm{I}}_\theta$, Θ^{I}, μ^{I}_Θ, $\mathrm{P}_{\mathrm{J},\mathrm{I}}$ and $\mathrm{P}^*_{\mathrm{J},\mathrm{I}}$ are the respective objects corresponding to X^{I} and X^{J}. For the observation variables X^{I}, on the other hand, I will always denote the actual set of axes.

Example 32 (Parametric family of Gaussian processes)**.** Instead of an individual Gaussian process measure as described in example 29, consider a parametric family of Gaussian processes: X^{E} once more takes values in $\Omega^{\mathrm{E}} = \mathbb{R}^{\mathbb{Z}}$. Let Θ^{E} be a random variable $(\Lambda, \mathcal{A}) \to (\Omega^{\mathrm{E}}_\theta, \mathcal{B}^{\mathrm{E}}_\theta)$ assuming as values pairs $(m, \Sigma) \in \mathbb{R}^{\mathbb{Z}} \times \mathbb{R}^{\mathbb{Z}\times\mathbb{Z}}$, such that the restriction of $\mu^{\mathrm{E}}_\Theta := \Theta^{\mathrm{E}}(\mathbb{P})$ to $\mathbb{R}^{\mathbb{Z}\times\mathbb{Z}}$ concentrates on the subset of positive definite operators. In customary terminology, m represents the mean function and Σ the covariance operator of a given Gaussian process measure. A parametric family of Gaussian processes is then defined as the conditional $\mu^{\mathrm{E}}(X^{\mathrm{E}}|\Theta^{\mathrm{E}})$. Since $\mathbb{Z}$ is countable and $\mathbb{R}$ Polish, $\mathbb{R}^{\mathbb{Z}}$ is Polish, and $(\Omega^{\mathrm{E}} = \mathbb{R}^{\mathbb{Z}}, \mathcal{B}^{\mathrm{E}})$ is standard Borel, hence the conditional distribution $\mu^{\mathrm{E}}(X^{\mathrm{E}}|\Theta^{\mathrm{E}})$ is regular (that is, it has a regular version).

Now consider the $\mathrm{P}^*_{\mathrm{E},\mathrm{I}}$-projections of the parametric family. The σ-subalgebras $\mathcal{C}^{\mathrm{I}}$ sufficient for each X^{I} can be identified explicitly. Both $\mathbb{R}^{\mathbb{Z}}$ and $\mathbb{R}^{\mathbb{Z}\times\mathbb{Z}}$ have a natural product structure. The marginal variables X^{I} follow Gaussian distributions, and for each finite $I \subset \mathbb{Z}$, a Gaussian on Ω^{I}_x can be completely specified by a parameter value in $\mathbb{R}^{\mathrm{I}} \times \mathbb{R}^{\mathrm{I}\times\mathrm{I}}$. We therefore define the projector $\mathrm{P}_{\mathrm{E},\mathrm{I}}$ on parameter space, with notation overloaded in the sense of Rem. 31, as the projector $\mathbb{R}^{\mathbb{Z}} \times \mathbb{R}^{\mathbb{Z}\times\mathbb{Z}} \to \mathbb{R}^{\mathrm{I}} \times \mathbb{R}^{\mathrm{I}\times\mathrm{I}}$. The parameter variables of the marginals are then defined as $\Theta^{\mathrm{I}} := \mathrm{P}_{\mathrm{E},\mathrm{I}}\Theta^{\mathrm{I}}$, and hence random variables $(\Lambda, \mathcal{A}) \to (\Omega^{\mathrm{I}}_\theta, \mathcal{B}^{\mathrm{I}}_\theta)$. Then obviously, $\sigma(\Theta^{\mathrm{I}}) = \Theta^{\mathrm{I},-1}(\mathcal{B}^{\mathrm{I}}_\theta) \subset \Theta^{\mathrm{J},-1}(\mathcal{B}^{\mathrm{J}}_\theta) = \sigma(\Theta^{\mathrm{J}})$ for $I \subset J$. Therefore, the parametric model $\mathrm{P}^*_{\mathrm{E},\mathrm{I}}\mu^{\mathrm{E}}(X^{\mathrm{E}}|\Theta^{\mathrm{E}})$ is the $|I|$-dimensional Gaussian family on $\mathbb{R}^{\mathrm{I}}$.

5.4 Construction Results

This section will make precise the results outlined in Sec. 5.2. We will first discuss the projective limit extension of regular conditional probabilities, and then apply these to Bayesian systems. The Bayesian limit systems can generally be guaranteed to be regular if the dimension of the extension space is countable. In the particular case of closed-form conjugate posteriors, a regular conditional projective limit can be guaranteed even for uncountable dimensions.

Remark 33 (Implications of Pitman-Koopman theory)**.** Before we turn to construction results and their proof, we should point out an immediate but important consequence of Pitman-Koopman theory (Sec. 2.2.4). The existence of sufficient statistics and conjugate priors is, at least for dominated and reasonably smooth models, effectively equivalent: Sampling models admitting sufficient statistics are, by the Pitman-Koopman theorem, exponential family models and admit a conjugate prior. Conversely, by La. 15, conjugate models admit sufficient statistics – in fact, conjugate priors are defined by means of sufficient statistics by several authors, including Raiffa and Schlaifer (1961) and Bernardo and Smith (1994). It is straightforward to show that all marginals of a conjugate infinite-dimensional Bayesian model are conjugate. The latter is true whenever there is a well-defined notion of a projector, and even the general topological assumptions of the Kolmogorov extension theorem are not required. Consequently, when a stochastic process Bayesian model is defined for exchangeable observations, existence of a conjugate posterior is guaranteed in general only if the model is constructed from marginals that are exponential families, or mixtures thereof. Though such exponential families may be arbitrarily complicated, and may involve combinations of different standard exponential families models (such as for Pólya tree models), there is simply no way for a non-conjugate system of marginals to extend to a conjugate projective limit. Conjugate process models are essentially restricted to projective limits of exponential families.

5.4.1 Infinite extension of conditional probabilities

The extension of regular conditional probabilities is summarized by Prop. 36 below. To extend the conditional, we have to take a projective limit with respect to both the sample variable and the parameter variable (or the σ-algebra it generates). The following two lemmas each cover one of these cases: We first consider the limit of a conditional distribution on a fixed sample space, but conditional with respect to consecutively finer σ-algebras. The second lemma keeps the σ-algebra representing the condition fixed, and extends the sample variable. Both are then combined, yielding Prop. 36.

Lemma 34. *Let E be an infinite set, $X : (\Lambda, \mathcal{A}) \rightarrow (\Omega_x, \mathcal{A}_x)$ be a random variable and $\{\mathcal{C}^I | I \in E^*\}$ a system of σ-subalgebras of $\mathcal{A}$, such that $\mathcal{C}^I \subset \mathcal{C}^J$*

whenever $I \subset J$. Define

$$\mathcal{C}^{E} := \sigma\Big(\bigcup_{I \in E^*} \mathcal{C}^{I}\Big)\,. \tag{5.4.1}$$

(1.) *The conditional probability $\mu(\,.\,|\mathcal{C}^E)$ is uniquely determined, up to equivalence, by the conditional probabilities $\mu(\,.\,|\mathcal{C}^I)$ for $I \in E^*$.*
(2.) *If $(\Omega_x, \mathcal{A}_x)$ is Borel, the collection uniquely defines a Markov kernel K_μ. If $\tilde{X}$ is any random variable with the $\mu(\,.\,|\mathcal{C}^I)$ as versions of its conditional probabilities for all $I \in E^*$, then K_μ is a regular conditional probability for $\tilde{X}$.*

Proof. **(1.)** Choose an arbitrary nested sequence $I_1 \subset I_2 \subset \dots$ of finite subsets of E. Then $j \mapsto \mathcal{C}^{I_j}$ is a filtration. For all $A \in \mathcal{A}_x$, define

$$\mu(A|\mathcal{C}^{\mathrm{I}}) = \mathbb{E}\left[\mathbb{I}_{X^{-1}(A)}|\mathcal{C}^{\mathrm{I}}\right] =: Z^{\mathrm{I}}_A\,. \tag{5.4.2}$$

Then by theorem 46, $(Z^{\mathrm{I}_j}, \mathcal{C}^{\mathrm{I}_j})$ forms a uniformly integrable martingale. By theorem 47, there exists one and only one $\mathcal{C}^{\mathrm{E}}$-measurable random variable Z^{E}_A such that $Z^{\mathrm{I}}_A = \mathbb{E}\,[Z^{\mathrm{E}}_A|\mathcal{C}^{\mathrm{I}}]$ for all $j \in \mathbb{N}$. Since $\mathcal{C}^{\mathrm{E}} \subset \mathcal{A}$, the conditional probability $\mu(A|\mathcal{C}^{\mathrm{E}})$ exists for each A and is a.s.-unique. Moreover, $\mathbb{E}\left[|\mathbb{I}_{X^{-1}(A)}|\right] < \infty$ for any A, and therefore $Z^{\mathrm{I}}_A = \mathbb{E}\,[\mu(A|\mathcal{C}^{\mathrm{E}})|\mathcal{C}^{\mathrm{I}}]$ by the law of total probability. Hence by uniqueness of Z^{E}_A,

$$Z^{\mathrm{E}}_A = \mu(A|\mathcal{C}^{\mathrm{E}}) \qquad\qquad a.e. \tag{5.4.3}$$

(2.) If $(\Omega_x, \mathcal{A}_x)$ is Borel, $\mathcal{A}_x$ is countably generated. Then so is $\mathcal{C}^{\mathrm{E}} \subset \mathcal{A}_x$, by some countable system $\mathcal{G}$. Choose a version $\mu(\,.\,|\mathcal{C}^{\mathrm{E}})(\omega)$ for each ω and define

$$P(A, \omega) := \mu(\,.\,|\mathcal{C}^{\mathrm{E}})(\omega) \qquad\qquad \text{for all } A \in \mathcal{G}\,. \tag{5.4.4}$$

By lemma 45, the extension of P from $\mathcal{G}$ to $\mathcal{C}^{\mathrm{E}}$ is a Markov kernel on $\mathcal{C}^{\mathrm{E}}$. What remains to be shown is that it actually coincides with the conditional probability defined by the limit process. But $P(A, \omega)$ is, by its definition in (5.4.4), a version of $\mu(A|\mathcal{C}^{\mathrm{E}})$ for all $A \in \mathcal{G}$, and so

$$\forall C \in \mathcal{C}^{\mathrm{E}}: \qquad\qquad \int_C P(A, \omega) d\mathbb{P}(\omega) = \mathbb{P}(C \cap \{X \in A\})\,. \tag{5.4.5}$$

That is, $\int_C P(\,.\,, \omega) d\mathbb{P}(\omega)$ and $\mathbb{P}(C \cap \{X \in \,.\,\})$ coincide on the generator, which is an algebra, and are both finite measures (because P is Markov). Then, by the uniqueness theorem for measures, both coincide on $\sigma(\mathcal{G}) = \mathcal{C}^{\mathrm{E}}$, and the Markov kernel defined by extension of P to $\mathcal{C}^{\mathrm{E}}$ is a conditional probability for X. □

Lemma 35. *Let $\{\mu^I | I \in E^*\}$ be a projective family, with μ^E its projective limit, and $\mathcal{C} \subset \mathcal{A}$ a sub-σ-algebra. Let $\mu^I(\,.\,|\mathcal{C})$ be regular conditional probabilities of the μ^I. Then if E is countable, there is a $\mathbb{P}$-null set N such that:*

1. *For all $\omega \in \Lambda \setminus N$, the family of measures $\{\mu^I(\,.\,|\mathcal{C})(\omega) | I \in E^*\} := \mathcal{M}_{\omega,\mathcal{C}}$ is projective.*
2. *Let ν be an arbitrary probability measure on $(\Omega^E, \mathcal{B}^E)$. The Markov kernel $\mu^E(\,.\,|\mathcal{C})$ defined by*

$$\begin{aligned} \mu^E(\,.\,|\mathcal{C})(\omega) &:= \operatorname{proj\,lim} \mathcal{M}_{\omega,\mathcal{C}} && \omega \in M \setminus N \\ \mu^E(\,.\,|\mathcal{C})(\omega) &:= \nu && \omega \in N \end{aligned} \tag{5.4.6}$$

is a regular conditional probability of μ^E.

Proof. Let $I \subset J \in E^*$. Then for any $B^{\mathrm{I}} \in \mathcal{B}^{\mathrm{I}}$, $\mu^{\mathrm{J}}(\mathrm{R}_{\mathrm{J,I}} B^{\mathrm{I}} | \mathcal{C})$ is a version of

$$\mathbb{P}\{X^{\mathrm{J}} \in \mathrm{R}_{\mathrm{J,I}} B^{\mathrm{I}} | \mathcal{C}\} = \mathbb{P}(X^{\mathrm{J},-1}(\mathrm{R}_{\mathrm{J,I}} B^{\mathrm{I}}) | \mathcal{C}) = \mathbb{P}(X^{\mathrm{I},-1}(B^{\mathrm{I}}) | \mathcal{C}) = \mathbb{P}\{X^{\mathrm{I}} \in B^{\mathrm{I}} | \mathcal{C}\} \,, \tag{5.4.7}$$

of which in turn $\mu^{\mathrm{I}}(B^{\mathrm{I}} | \mathcal{C})$ is a version. Since B^{J} has a countable generator, there exists an $\mathbb{P}$-null set $N^{\mathrm{IJ}} \subset \Lambda$, such that:

$$\forall \omega \in \Lambda \setminus N^{\mathrm{IJ}} : \quad \mu^{\mathrm{J}}(\mathrm{R}_{\mathrm{J,I}}\,.\,|\mathcal{C})(\omega) = \mu^{\mathrm{I}}(\,.\,|\mathcal{C})(\omega) \,. \tag{5.4.8}$$

Let $N^* := \bigcup_{I \subset J \in E^*} N^{\mathrm{IJ}}$. Since E is countable, so is E^*, and N^* is a null set. Hence for each $\omega \in \Lambda \setminus N^*$, the family $\mathcal{M}_{\omega,\mathcal{C}}$ is projective.

Since E is countable, the product space Ω^{E} is Polish, hence μ^{E} has regular conditional probabilities $\mu^{\mathrm{E}}(\,.\,|\mathcal{C})$, and these differ only on an $\mathbb{P}$-null set N^{E}. Let $N := N^* \cup N^{\mathrm{E}}$, for which again $\mathbb{P}(N) = 0$. Since by construction, (5.4.7) applies also to $\mu^{\mathrm{E}}(\,.\,|\mathcal{C})$, the marginal measures of $\mu^{\mathrm{E}}(\,.\,|\mathcal{C})(\omega)$ are $\mu^{\mathrm{I}}(\,.\,|\mathcal{C})(\omega)$ for all $\omega \in \Lambda \setminus N$. □

To guarantee regularity of the limit conditional, La. 34 relies on a Borel structure of the limit space, and La. 35 on countability of the index set E. Countability of E again implies a Borel structure on the limit space (since countable products of Polish spaces are Polish). If the Borel assumption is dropped in the first lemma, we still obtain a conditional probability, but it need not be regular. Since a simple, non-regular conditional may be perfectly satisfactory for some problems, it is interesting to ask what happens if the countability is not demanded in La. 35. But the problem here is the set N, because for any pair of subspaces I, J, there is a null set of exceptions ω for which the conditional is not projective (the sets denoted

N^{IJ} in the proof above). If E is not countable, these may aggregate into a non-null set – in principle, the union N^* may be the entire space. In that case, the limit conditional, which is only determined by the marginal on $\complement N$, is not determined anywhere. In short, the limit conditional is reasonably well-specified by a martingale argument if the observation variable is fixed, but generalizing beyond the countably-dimensional case seems to be difficult for conditional measures on finite-dimensional subspaces. At the very least, the generalization would require more sophisticated methods than the somewhat simplistic approach taken here. On the other hand, there is one notable exception: If the sets of exceptions are empty, which they are for standard parametric models defined in terms of parametric densities, a case considered in Prop. 43 below.

Combining the two results above gives the following proposition, a rough analogue of the extension theorem for the conditional case.

Proposition 36. *Let E be a countable set, possibly infinite, and Ω^E a product space with Polish components Ω. Let $\{\Omega^I | I \in E^*\}$ be the system of all finite-dimensional subspaces. Let $\{\mathcal{C}^I | I \in E^*\}$ be a system of σ-subalgebras of $\mathcal{A}$, satisfying $\mathcal{C}^I \subset \mathcal{C}^J$ whenever $I \subset J$. Let $X^I : (\Lambda, \mathcal{A}) \to (\Omega^I, \mathcal{B}^I)$ be random variables with values on the subspaces. Assume that the regular conditional probabilities $\mu^I(X^I | \mathcal{C}^I)$ are projective in the sense*

$$\forall I \subset J \in E^* : \qquad (\mathrm{P}^*_{\mathrm{J,I}} \mu^J)(X^I | \mathcal{C}^J) = \mu^I(X^I | \mathcal{C}^I) \; . \qquad (5.4.9)$$

Then there exists a Markov kernel $\mu^E(\, . \, | \mathcal{C}^E)(\, . \,)$ such that:

1. *For all $I \in E^*$, and all $A^I \in \mathcal{B}^I$:*
$$\mathrm{P}^*_{\mathrm{E,I}} \mu^E(A^I | \mathcal{C}^E) = \mu^I(A^I | \mathcal{C}^I) \qquad \mathbb{P} - a.e. \qquad (5.4.10)$$
2. *There is a $\mathbb{P}$-null set N such that $\mu^E(\, . \, | \mathcal{C}^E)(\, . \,)(\omega)$ is unique for all $\omega \in \complement N$.*
3. *If $\tilde{X}$ is any random variable on $(\Omega^E, \mathcal{B}^E)$ with $(\mathrm{P}^*_{\mathrm{E,I}} \tilde{\mu}^E)(A^I | \mathcal{C}^E) = \mu^I(A^I | \mathcal{C}^I)$ a.e., then $\mu(\, . \, | \mathcal{C}^E)$ is a regular conditional probability of $\tilde{X}^E$.*

For Bayesian models, the extension of parametric models is of particular interest. That is, an infinite-dimensional space Ω_x^{E} with finite-dimensional subspaces Ω_x^{I} is defined as above, and a parametric model $\mu^{\mathrm{I}}(X^{\mathrm{I}} | \Theta^{\mathrm{I}})$ is defined on each subspace. The parameter variable Θ^{I} is assumed to take values in some space $\Omega_\theta^{\mathrm{I}}$. This does not necessarily imply that the dimension of a parameter θ^{I}, when regarded as a vectorial quantity, is the same as that of the random variable X^{I}. If, for example, the values x^{I} are d-dimensional vectors (with $d = |I|$), the corresponding parameter values θ^{I} may be $d \times d$-matrices, such as a Gaussian covariance parameter. θ^{I} may also be a tuple of quantities, e. g. a d-vector and $d \times d$-matrix in the fully parameterized Gaussian case. For these two examples, Ω^{I} would be $\mathbb{R}^d$, and suitable parameter spaces would be $\mathbb{R}^{d\times d}$ in the former and $\mathbb{R}^d \times \mathbb{R}^{d\times d}$ in the latter case. Therefore, we need to make an important distinction between Ω_x^{I} and $\Omega_\theta^{\mathrm{I}}$: While the sample space Ω_x^{I} is the product $\prod_{i\in I}\Omega_x$, the I-notation $\Omega_\theta^{\mathrm{I}}$ for the parameter space only indicates that it corresponds to Ω_x^{I}, whereas the actual index set may differ from I. The only general assumption we will have to make is that $\Omega_\theta^{\mathrm{I}} \subset \Omega_\theta^{\mathrm{J}}$ whenever $I \subset J$, that is, higher-dimensional observations have higher-dimensional parameters (in a non-strict sense).

Remark 37 (Extension of parametric families)**.** To apply proposition 36 to the extension of a system $\{\mu^{\mathrm{I}}(X^{\mathrm{I}}|\Theta^{\mathrm{I}})\}$ of parametric families, choose the system such that the parameter variables are projective: We overload notation (cf. Rem. 31) and write $\mathrm{P}_{\mathrm{J,I}}$ for the projector from $\Omega_\theta^{\mathrm{J}}$ to $\Omega_\theta^{\mathrm{I}}$. The index sets I, J are symbolic in the sense explained above, but since the spaces have product structure, the projector is well-defined for any $I \subset J$. Then the parameter variables are projective if $\mathrm{P}_{\mathrm{J,I}}\Theta^{\mathrm{J}} = \Theta^{\mathrm{I}}$. Now the σ-subalgebras $\mathcal{C}^{\mathrm{I}}$ in the proposition are defined as those generated by the parameters variables, $\mathcal{C}^{\mathrm{I}} := \sigma(\Theta^{\mathrm{I}})$. These σ-algebras are nested as required by the propositions, since if $I \subset J$ and $B^{\mathrm{I}} \in \mathcal{B}_\theta^{\mathrm{I}}$, then $\sigma(\Theta^{\mathrm{I}}) = \Theta^{\mathrm{I},-1}(B^{\mathrm{I}}) = \Theta^{\mathrm{J},-1}(\mathrm{R}_{\mathrm{J,I}}B^{\mathrm{I}})$. As projective random variables on Polish product spaces, the parameters have an infinite-dimensional projective limit Θ^{E} by Kolmogorov's extension theorem. The σ-algebra generated by this limit variable has to coincide with the limit $\mathcal{C}^{\mathrm{E}}$ defined in 5.4.1 for the proposition to be applicable. This is indeed the case. The generated σ-algebra is $\sigma(\Theta^{\mathrm{E}}) = \Theta^{\mathrm{E},-1}(\mathcal{B}^{\mathrm{E}})$, where $\mathcal{B}^{\mathrm{E}}$ is the product σ-algebra on Ω^{E}. Since $\mathcal{B}^{\mathrm{E}}$ is generated by the cylinder sets, and therefore:

$$\begin{aligned}\Theta^{\mathrm{E},-1}(\mathcal{B}^{\mathrm{E}}) &= \Theta^{\mathrm{E},-1}\Big(\sigma\big(\bigcup_{I\in E^*}\mathrm{R}_{\mathrm{E,I}}\mathcal{B}^{\mathrm{I}}\big)\Big) = \sigma\big(\bigcup_{I\in E^*}\Theta^{\mathrm{E},-1}(\mathrm{R}_{\mathrm{E,I}}\mathcal{B}^{\mathrm{I}})\big)\\ &= \sigma\big(\bigcup_{I\in E^*}\Theta^{\mathrm{I},-1}(\mathcal{B}^{\mathrm{I}})\big) = \sigma\big(\bigcup_{I\in E^*}\mathcal{C}^{\mathrm{I}}\big) = \mathcal{C}^{\mathrm{E}}\end{aligned} \tag{5.4.11}$$

Then proposition 36 states that the family $\{\mu^{\mathrm{I}}(X^{\mathrm{I}}|\Theta^{\mathrm{I}})|I \in E^*\}$ uniquely (up to an $\mathbb{P}$-null set) defines a regular conditional probability $\mu^{\mathrm{E}}(\,.\,|\Theta^{\mathrm{E}})$ on Ω_x^{E} such that for any $\omega \in \Lambda$, the $\mathrm{P}^*_{\mathrm{E,I}}$-marginals of $\mu^{\mathrm{E}}(\,.\,|\Theta^{\mathrm{E}})$ are $\mu^{\mathrm{I}}(\,.\,|\Theta^{\mathrm{I}})$.

5.4.2 Bayesian Extension and Sufficiency

The results above immediately apply to some aspects of Bayesian models, as summarized by the following two corollaries. If we can define unique limits (up to equivalence) of conditional probabilities, then we can define limits of complete Bayesian equations, which are conjugate if and only if the finite-dimensional systems used in the definition are conjugate. In the countably-dimensional case, sufficient statistics of the finite-dimensional projective families determine a sufficient statistic of the infinite-dimensional limit. If the finite-dimensional marginal components are conjugate exponential families, we can in fact guarantee a regular Bayesian equation with infinite-dimensional sufficient statistic and conjugate posterior, even if the dimension is uncountable.

One consequence in the countable case is that the projective limits of marginals that satisfy a Bayesian equation on each finite-dimensional subspace again satisfy a Bayesian equation.

Corollary 38 (Projective limits of Bayesian equations). *Let E be countable. Let $\{\mu^I(X^I|\Theta^I)|I \in E^*\}$ be a family of finite-dimensional (regular conditional) parametric models, each with a prior $\mu^I_\Theta(\Theta)$. Denote by $\mu^I(\Theta^I|X^I)$ the corresponding posteriors, each assumed to be regular. If (1) the family $\{\mu^I_\Theta(\Theta)|I \in E^*\}$ is projective and if (2) either $\{\mu^I(X^I|\Theta^I)|I \in E^*\}$ or $\{\mu^I(\Theta^I|X^I)|I \in E^*\}$ is projective with respect to $\mathrm{P}^*_{\mathrm{J,I}}$, then there is a uniquely defined Bayesian model on Ω^E_x, Ω^E_θ, with the given finite-dimensional models as its marginals.*

That is, the prior-posterior relation is preserved under the projective limit. The same is true for conjugacy.

Corollary 39 (Conjugate projective limits). *Consider a family $\{\mu^I(X^I|\Theta^I)|I \in E^*\}$ of parametric models as in Cor. 38. Assume that for each, a family of priors $\mathcal{N}^I = \{\mu^I_{\Theta,y}|y \in \mathcal{Y}^I\}$ is given. Then if the finite-dimensional Bayesian systems are conjugate for each $I \in E^*$, the family infinite-dimensional extensions is conjugate, with respect to the family $\mathcal{N}^E$ defined as follows: $\mathcal{N}^E$ is the set of all measures μ^E_Θ such that:*

$$\forall I \in E^* \exists y^I \in \mathcal{Y}^I : \qquad \mu^E_\Theta = \operatorname{proj\,lim}\{\mu^I_{\Theta,y^I}|I \in E^*\} \,. \tag{5.4.12}$$

Proposition 40 (Extension of sufficient statistics). *Let X^E be a random variable with values in a product Ω^E_x of Polish spaces, with σ-algebra $\mathcal{B}(\Omega^E_x)$. Let $\mu^E(X^E|\Theta^E)$ be a parametric model for X^E, and $\mathcal{N}^E = \{\mu^E_{\Theta,y}|y \in \mathcal{Y}\}$ a family of priors, indexed by some non-empty set $\mathcal{Y}$. Let further Ω^E_s be a product space and $\{s^I : (\Omega^I_x, \mathcal{B}^I_x) \to (\Omega^I_s, \mathcal{B}^I_s)\}$ a family of measurable maps. Finally, let $\mu^I(X^I|\Theta^I) := \mathrm{P}^*_{\mathrm{E,I}}\mu^E(X^E|\Theta^E)$ be the marginals of the parametric model. Assume that:*

1. *E is countable.*
2. *For each I, s^I is sufficient for Θ^I with respect to the marginal $\mu^I(X^I|\Theta^I)$.*

Then any measurable map $s^E : \Omega_x^E \to \Omega_s^E$ *satisfying*

$$\forall I \in E^*, A^I \in \mathcal{B}_\theta^I : \qquad s^{E,-1}(\mathrm{R}_{\mathrm{E,I}} A^I) = \mathrm{R}_{\mathrm{E,I}} s^{I,-1}(A^I) \tag{5.4.13}$$

is sufficient for Θ^E.

Proof. The posteriors of Θ^{E} under observation of X^{E} are $\mu^{\mathrm{E}}_{\Theta,y}(\Theta^{\mathrm{E}}|\sigma(X^{\mathrm{E}}))$, and their marginals for finite I are $\mu^{\mathrm{I}}_{\Theta,y}(\Theta^{\mathrm{I}}|\sigma(X^{\mathrm{I}}))$, respectively. Sufficiency of s^{I} is equivalent to

$$\mu^{\mathrm{I}}_{\Theta,y}(\Theta^{\mathrm{I}}|\sigma(X^{\mathrm{I}})) = \mu^{\mathrm{I}}_{\Theta,y}(\Theta^{\mathrm{I}}|\sigma(X^{\mathrm{I}} \circ s^{\mathrm{I}})) \qquad \text{a.s.} \tag{5.4.14}$$

Proposition 36 specifies the projective limits of both sides of the equation. For the left-hand side, the limit is

$$\mu^{\mathrm{E}}_{\Theta,y}\Big(\Theta^{\mathrm{E}}\Big|\sigma\big(\bigcup_{I\in E^*}\sigma(X^{\mathrm{I}})\big)\Big) = \mu^{\mathrm{E}}_{\Theta,y}(\Theta^{\mathrm{E}}|\sigma(X^{\mathrm{E},-1}(\mathcal{Z}_x^{\mathrm{E}}))) = \mu^{\mathrm{E}}_{\Theta,y}(\Theta^{\mathrm{E}}|\sigma(X^{\mathrm{E}}))\ , \tag{5.4.15}$$

which is just the posterior for Θ^{E}. The limit of the right-hand side is

$$\mu^{\mathrm{E}}_{\Theta,y}\Big(\Theta^{\mathrm{E}}\Big|\sigma\big(\bigcup_{I\in E^*}\sigma(X^{\mathrm{I}} \circ s^{\mathrm{I}})\big)\Big)\ . \tag{5.4.16}$$

Both must be identical a. e. by uniqueness, so s^{E} will indeed determine the posterior (and hence be sufficient), provided that $\sigma\big(\bigcup_{I\in E^*}\sigma(X^{\mathrm{I}} \circ s^{\mathrm{I}})\big) = \sigma(X^{\mathrm{E}} \circ s^{\mathrm{E}})$. But the latter is true, because

$$\begin{aligned}\sigma(X^{\mathrm{E}} \circ s^{\mathrm{E}}) &= X^{\mathrm{E},-1}(s^{\mathrm{E},-1}(\mathcal{B}_s^{\mathrm{E}})) = X^{\mathrm{E},-1}\Big(s^{\mathrm{E},-1}\Big(\sigma\big(\bigcup_{I\in E^*}\mathrm{R}_{\mathrm{E,I}}\mathcal{B}_s^{\mathrm{I}}\big)\Big)\Big)\\ &= \sigma\Big(\bigcup_{I\in E^*} X^{\mathrm{E},-1}\big(\mathrm{R}_{\mathrm{E,I}} s^{\mathrm{I},-1}(\mathcal{B}_s^{\mathrm{I}})\big)\Big) = \sigma\Big(\bigcup_{I\in E^*} X^{\mathrm{I},-1}(s^{\mathrm{I},-1}(\mathcal{B}_s^{\mathrm{I}}))\Big)\\ &= \sigma\Big(\bigcup_{I\in E^*}\sigma(X^{\mathrm{I}} \circ s^{\mathrm{I}})\Big)\ .\end{aligned} \tag{5.4.17}$$

□

Remark 41 (Sufficient condition for s^{E})**.** A sufficient condition for the Bayesian sufficiency of the map s^{E} is given by Eq. (5.4.13). An alternative sufficient condition, which is less general, but easier to verify for most functions, is the following:

$$\forall I \in E^*, x^{\mathrm{E}} \in \Omega_x^{\mathrm{E}} : \qquad \mathrm{P}_{\mathrm{E,I}} s^{\mathrm{E}}(x^{\mathrm{E}}) = s^{\mathrm{I}}(\mathrm{P}_{\mathrm{E,I}} x^{\mathrm{E}})\ . \tag{5.4.18}$$

Any mapping s^{E} satisfying this condition also satisfies (5.4.13): For any $\theta^{\text{I}} \in \Omega_{\theta}^{\text{I}}$,

$$\begin{aligned}
\text{R}_{\text{E,I}} s^{\text{I},-1}(\theta^{\text{I}}) &= \text{R}_{\text{E,I}}\{x^{\text{I}} | s^{\text{I}}(x^{\text{I}}) = \theta^{\text{I}}\} = \{x^{\text{E}} | \text{P}_{\text{E,I}} x^{\text{E}} = x^{\text{I}} \wedge s^{\text{I}}(x^{\text{I}}) = \theta^{\text{I}}\} \\
&= \{x^{\text{E}} | s^{\text{I}}(\text{P}_{\text{E,I}} x^{\text{E}}) = \theta^{\text{I}}\} = \{x^{\text{E}} | \text{P}_{\text{E,I}} s^{\text{E}}(x^{\text{E}}) = \theta^{\text{I}}\} \\
&= \{x^{\text{E}} | s^{\text{E}}(x^{\text{E}}) = \theta^{\text{E}} \wedge \text{P}_{\text{E,I}}(\theta^{\text{E}}) = \theta^{\text{I}}\} = s^{\text{E},-1}\big(\{\theta^{\text{E}} | \text{P}_{\text{E,I}}(\theta^{\text{E}}) = \theta^{\text{I}}\}\big) \\
&= s^{\text{E},-1}(\text{R}_{\text{E,I}}\{\theta^{\text{I}}\}) \; .
\end{aligned} \tag{5.4.19}$$

This reasoning extends immediately from points θ^{I} to sets A^{I}. Note that Eq. (5.4.18), which expresses a relation between the infinite-dimensional and finite-dimensional case, implies an analogous relation between pairs of finite-dimensional sufficient statistics.

$$\forall I \subset J \in E^*, x^{\text{J}} \in \Omega_x^{\text{J}} : \qquad \text{P}_{\text{J,I}} s^{\text{J}}(x^{\text{J}}) = s^{\text{I}}(\text{P}_{\text{J,I}} x^{\text{J}}) \; . \tag{5.4.20}$$

If such an s^{E} exists, it is unique, since for any x^{E}, the image $s^{\text{E}}(x^{\text{E}})$ is determined pointwise by $(s^{\text{E}}(x^{\text{E}}))_i = s^{\{i\}}(\{i\})$.

Remark 42 (Generalization of Proposition 40)**.** Proposition 40 assumes a Borel product structure on the range of the sufficient statistics s^{I}. This assumption mainly serves to simplify the statement of the result (rather than the proof), and is straightforward to generalize. As a look at the proof shows, actual projectiveness of the form $X^{\text{J},-1}(\text{R}_{\text{J,I}} A^{\text{I}}) = X^{\text{I},-1}(A^{\text{I}})$ is required only for the observation variables X^{I}. For the sufficient statistic, the condition can be relaxed by assuming that all s^{I} map into some common space Ω_s, requiring only that $\sigma(s^{\text{I}}) \subset \sigma(s^{\text{J}})$ whenever $I \subset J$. We then need some operation $\tilde{\text{R}}_{\text{J,I}}$ to substitute for the preimage operation $\text{R}_{\text{J,I}}$. The operation must be "dual" to $\text{R}_{\text{J,I}}$ with respect to the mappings s^{I} in the following sense: $\tilde{\text{R}}_{\text{J,I}}$ is a mapping $\Omega_s \rightarrow \Omega_s$, and for all $I \subset J$,

$$s^{\text{J}}(\tilde{\text{R}}_{\text{J,I}} A^{\text{I}}) = \text{R}_{\text{J,I}} s^{\text{I}}(A^{\text{I}}) \; . \tag{5.4.21}$$

Additionally, the σ-algebras in the range of the statistics s^{I}, assumed to be product Borel algebras $\mathcal{B}_s^{\text{I}}$ in the statement of the proposition above, can be substituted by any other σ-algebras $\mathcal{C}_s^{\text{I}}$, which have to satisfy only one requirement: For $I \subset J$, the "dual" preimages of sets in $\mathcal{C}_s^{\text{I}}$ must be contained in $\mathcal{C}_s^{\text{I}}$. Presumably, most such generalizations will result in a projective Polish structure on Ω_θ one way or the other, but they generalize the result for example to the case of parameter spaces that have a different dimension then the observation space, such as the Gaussian vector-matrix examples discussed in the previous section. It is also interesting to note that the topological requirement of Polish product structures, essential to the Kolmogorov construction, is not actually required on the range of the sufficient statistics.

Finally, for the arguably most interesting case of marginals (the exponential family case), a *regular* Bayesian limit system is available even in the uncountably-dimensional case.

Common density models, including in particular all exponential family models, define a measure for *each* parameter in their parameter set, rather than for almost all parameters. There are no non-empty null sets of exceptions which may aggregate into non-null sets in the limit of uncountable dimensions. However, this observation extends from prior and likelihood to the posterior only if the posterior has a closed-form density, which is just the case for conjugate exponential models. In this case, the relation between projective limit models and finite-dimensional marginals is one-to-one, and the diagram above is fully commutative.

$$\begin{tikzcd}
\mu^{\mathrm{E}}(\Theta^{\mathrm{E}}|X^{\mathrm{E}},Y^{\mathrm{E}}) \arrow[d, shift right, "\mathrm{P}_{\mathrm{E,I}}"'] & & \mu^{\mathrm{E}}_{\Theta}(\Theta^{\mathrm{E}}|Y^{\mathrm{E}}) \arrow[ll, "s^{\mathrm{E}}"', "x^{\mathrm{E}}_1,\ldots,x^{\mathrm{E}}_n"] \arrow[d, shift left, "\mathrm{P}_{\mathrm{E,I}}"] \\
\mu^{\mathrm{I}}(\Theta^{\mathrm{I}}|X^{\mathrm{I}},Y^{\mathrm{I}}) \arrow[u, shift right, "\underleftarrow{\lim}"'] & & \mu^{\mathrm{I}}_{\Theta}(\Theta^{\mathrm{I}}|Y^{\mathrm{I}}) \arrow[ll, "s^{\mathrm{I}}"', "x^{\mathrm{I}}_1,\ldots,x^{\mathrm{I}}_n"] \arrow[u, shift left, "\underleftarrow{\lim}"]
\end{tikzcd}$$

Proposition 43 (Projective limits of exponential family Bayesian models)**.** *Let E be an index set, possibly uncountable, and Ω_x, Ω_θ and Ω_y Polish spaces with Borel algebras. $\mathcal{T}^I \subset \Omega^I_\theta$ and $\mathcal{Y}^I \subset \Omega^I_y$ be measurable, open and convex for each I, and projective in the sense $\mathrm{P}_{\mathrm{J,I}}\mathcal{T}^J = \mathcal{T}^I$ and $\mathrm{P}_{\mathrm{J,I}}\mathcal{Y}^J = \mathcal{Y}^I$. Let $\{F^I(x^I|\theta^I)|I \in E^*, \theta^I \in \mathcal{T}^I)$ be a collection (over all I) of exponential family models, and $\{G^I(\theta^I|\lambda, y^I)|I \in E^*, y^I \in \mathcal{Y}^I\}$ a collection of priors naturally conjugate to the F^I. Let $\mu^I(X^I|\Theta^I = \theta^I)$ and $\mu^I_\Theta(\Theta^I|Y^I = (\lambda, y^I)$ be the measures defined by densities F^I and G^I with respect to suitable carriers. Define $\mathcal{T}^E$ as $\bigcap_{I\in E^*} \mathrm{R}_{\mathrm{E,I}}\mathcal{T}^I$, and $\mathcal{Y}^E$ accordingly. If, for each $a \in \mathbb{R}_+$, each $\theta^E \in \mathcal{T}^E$ and each $\mathbf{y}^E \in \mathcal{Y}^E$, the respective families $\{\mu^I(X^I|\Theta^I = \theta^I)|I \in E^*\}$ and $\{\mu^I_\Theta(\Theta^I|Y^I = (\lambda, y^I)|I \in E^*\}$ are projective, then:*

1. *The collections uniquely define a Bayesian model on the product spaces Ω^E_x, Ω^E_θ.*
2. *The parameter set of the sampling distribution $\mu^E(X^E|\Theta^E)$ is $\mathcal{T}^E$.*
3. *For each y^E in $\mathcal{Y}^E$, the prior $\mu^E(\mathcal{T}^E|Y^E = (\lambda, y^E)) = 1$.*
4. *If there exists a measurable function s^E such that $\mathrm{P}_{\mathrm{E,I}}s^E(x^E) = s^I(\mathrm{P}_{\mathrm{E,I}}x^E)$ for all $x^E \in \Omega^E_x$, the posterior given y^E under observations $x^E_1, \ldots, x^E_n$*

is

$$\mu_\Theta(\Theta^E|X_1 = x_1, \ldots, X_n = x_n) = \mu_\Theta\Big(\Theta^E\Big|Y^E = \big(\lambda{+}n, y{+}\sum_{i=1}^{n} s^E(x_i^E)\big)\Big) \tag{5.4.22}$$

5. *The prior, sampling model and posterior of the limit model are regular conditional probabilities.*

Proof. (1) follows for prior and likelihood from the projectiveness of the respective families. (See below for the posterior.)
(2) and (3): From the definition of $\mathcal{T}^{\text{E}}$ and $\mathcal{Y}^{\text{E}}$ follows that these sets parameterize those and only those measures with marginals parmaterized by $\mathcal{T}^{\text{I}}$ and $\mathcal{Y}^{\text{I}}$. Note that both $\mathcal{T}^{\text{E}}$ and $\mathcal{Y}^{\text{E}}$ are convex, since the cylinders of convex sets are convex. If $T \subset \Omega_\theta^{\text{E}}$ is any measurable set, then for all $I \in E^*$, $\mu_\Theta^{\text{I}}(\text{P}_{\text{E,I}}T|Y^{\text{I}} = (\lambda, y^{\text{I}})) \neq 0$ only if $(\text{P}_{\text{E,I}}T) \cap \mathcal{T}^{\text{I}} \neq \emptyset$, and so $\mu_\Theta^{\text{E}}(T|Y^{\text{E}} = (\lambda, y^{\text{E}})) \neq 0$ only if $T \cap \mathcal{T}^{\text{E}} \neq \emptyset$. Therefore, $\mu_\Theta^{\text{E}}(\mathcal{T}^{\text{E}}|Y^{\text{E}} = (\lambda, y^{\text{E}})) = 1$.
(4) For any admissible posterior of the limit system, its marginals must coincide with the posteriors of the marginal systems. The latter are uniquely determined by conjugacy, through the parameter update $(\lambda, y) \mapsto (\lambda + n, y + \sum s(x_i))$. Since the posterior is a measure for *each* parameter, and the posterior family is projective due to projectiveness of priors and sufficient statistics, Kolmogorov's theorem is applicable, and yields a uniquely defined measure for each set $x_1^{\text{E}}, \ldots, x_n^{\text{E}}$ of observations and each hyperparameter (λ, y^{E}).
(5) The unique determination of all component measures of the limit system for any value of the parameter by the extension theorem implies that all conditional limits are regular. □

The results comes with a caveat often neglected in the nonparametric Bayesian literature. The domain of the constructed measure μ^{E} is the product σ-algebra $\mathcal{B}^{\text{E}}$. As discussed in Rem. 24, if E is not countable, $\mathcal{B}^{\text{E}}$ does not contain singletons. Sample observations are singletons, and the concept of a Bayesian system which cannot measure singletons is somewhat questionable.

5.5 Examples

5.5.1 Dirichlet Process on $\mathbb{Q}$

The first example is, for the sake of familiarity, a Dirichlet process model. However, to keep as many as possible of the results discussed above applicable, we consider the countably-infinite case of a Dirichlet process generating random measures on the set $\mathbb{Q}$ of rational numbers. [1] The first construction step is the choice of suitable sets for Ω_x, Ω_θ and the index set E. The DP will generate random probability measures. If P is such a random measure, then Ω_θ will serve as its range, and E its domain. (The range is Ω_θ rather than Ω_x, because the DP represents the prior component of the Bayesian model). Choose $\Omega_x = \Omega_\theta = [0,1]$, which is Polish, and $\mathcal{B}_x = \mathcal{B}_\theta = \mathcal{B}([0,1])$. The finite-dimensional marginals of the infinite-dimensional prior are Dirichlet distributions μ_Θ^{I}, with densities of the form

$$G^{\mathrm{I}}(\theta^{\mathrm{I}}|\lambda, y^{\mathrm{I}}) = \frac{1}{Z_G(\theta^{\mathrm{I}})} \exp\bigl(\sum_i (\lambda y_i^{\mathrm{I}} - 1) \log(\theta_i^{\mathrm{I}})\bigr) \ . \tag{5.5.1}$$

The index i runs over $|I|$ different elements, but the index set I is more complicated than just the set of numbers up to $|I|$. Each of the Dirichlet distributions can be interpreted as follows: Assume that $\mathbb{Q}$ has been subdivided into a finite number of non-overlapping (measurable) sets, Q_i. In the parlance of histogram methods, the sets Q_i form a binning of $\mathbb{Q}$. Then I is the set of all Q_i, and a draw from G^{I} generates a finite probability distribution θ^{I}, which assigns to each set Q_i a probability θ_i^{I}. Apparently, for a given size of I, there will be one set I for each possible partition of $\mathbb{Q}$ into $|I|$ disjoint sets. Since $\mathbb{Q}$ is countable, the union of all such I is the set of all countable subsets of $\mathbb{Q}$ (including $\mathbb{Q}$ itself and the empty set). This union is precisely the Borel σ-algebra on $\mathbb{Q}$. Consider $\mathbb{Q}$ as a topological space with Borel algebra: Since $\mathbb{Q}$ is Hausdorff, $\mathcal{B}(\mathbb{Q})$ is generated by the compact subsets, which are just the finite sets in $\mathbb{Q}$. The σ-algebra generated by the finite sets of rationals contains sets generated by countable unions, and hence are countable, as well as their complements. Therefore, $\mathcal{B}(\mathbb{Q})$ is the power set $\mathcal{P}(\mathbb{Q})$. This implies that E should be chosen as $\mathcal{B}(\mathbb{Q})$. By restricting E^* from the usual set of finite subsets of E to those finite collections which form disjoint partitions, we have – in analogy to the construction of Ferguson (1973) on the real line – adapted the index sets to the definition of the Dirichlet distribution as a distribution on a partition. (Formally, the restriction can be justified by noting that the corresponding cylinder sets of the partitions still generate $\mathcal{B}(\mathbb{Q}) = E$.) The sampling distributions conjugate to G^{I} are multinomial measures (one observation). The random values drawn from the multinomial are vectors x^{I} of $|I|$ entries, with exactly one

[1] The construction on the rationals can, in fact, substitute for the construction on the real line, because the set of cumulative distribution functions on $\mathbb{R}$ admits a bijective and bimeasurable mapping to its own restriction to $\mathbb{Q}$ (see e. g. Ghosh and Ramamoorthi, 2002).

non-zero entry of value 1. The densities are

$$F^{\mathrm{I}}(x^{\mathrm{I}}|\theta^{\mathrm{I}}) \propto \exp\big(\sum_i x_i^{\mathrm{I}} \log(\theta_i^{\mathrm{I}})\big) \ . \tag{5.5.2}$$

The sufficient statistic s^{I} is the identity on Ω^{I}. Since F^{I} and G^{I} are naturally conjugate exponential family models for each I, the posterior under an observation x^{I} is of the form $G^{\mathrm{I}}(\theta^{\mathrm{I}}|\lambda + 1, y + x^{\mathrm{I}})$. It is straightforward to verify that the family μ^{I} of priors is projective if and only if $\lambda^{\mathrm{I}} = \lambda^{\mathrm{J}}$ and $\mathrm{P}_{\mathrm{J,I}} y^{\mathrm{J}} = y^{\mathrm{I}}$ for all $I \subset J$. Since s^{I} is an identity mapping and the posterior is conjugate, the family of posteriors is projective if and only if the priors are projective. Proposition 36 then guarantees the existence of an infinite-dimensional Bayesian equation, with its prior the projective limit of the priors, a Dirichlet process. The limit system is conjugate by Cor. 39, and since $s^{\mathrm{E}} := \mathrm{Id}_{\Omega_x^{\mathrm{E}}}$ is measurable and satisfies (5.4.13), it is a sufficient statistic for the infinite-dimensional system. The posterior of the limit system under observation an x^{E} is therefore obtained by substituting $(\lambda + 1, y^{\mathrm{E}} + x^{\mathrm{E}})$ for (λ, y) in the prior, which is precisely the analogue of the DP posterior as established by Ferguson (1973).

5.5.2 Exponential Family Product Models

Assume that on the system of finite-dimensional subspaces Ω^{I} of Ω^{E}, exponential family parametric models are defined as in Def. 9. Let h^{I} be the identity mapping for each $I \in E^*$, and assume the sufficient statistics are of the form $s^{\mathrm{I}} : \Omega_x^{\mathrm{I}} \to \mathcal{T}^{\mathrm{I}} \subset \Omega_\theta^{\mathrm{I}}$, where the parameter sets $\mathcal{T}^{\mathrm{I}}$ are convex, open, and consistent with projection, $\mathcal{T}^{\mathrm{I}} = \mathrm{P}_{\mathrm{J,I}} \mathcal{T}^{\mathrm{J}}$ for any $I \subset J$. With respect to the general discussion of sufficient statistics in Sec. 5.4.2 above, the ranges Ω_s^{I} of the sufficient statistics are chosen as the respective parameter spaces.

We will consider here only the case where spaces are real, with standard inner product on Ω_t^{I}, and the carrier measure is Lebesgue, denoted on both sample and parameter space by λ^{I} for I finite. The parametric models $\mu^{\mathrm{I}}(X^{\mathrm{I}}|\Theta^{\mathrm{I}})$ for the sampling distributions are specified by the densities

$$F^{\mathrm{I}}(x^{\mathrm{I}}|\theta^{\mathrm{I}}) = \frac{1}{Z_F^{\mathrm{I}}(\theta^{\mathrm{I}})} \exp(\langle s^{\mathrm{I}}(x^{\mathrm{I}})|\theta^{\mathrm{I}}\rangle) \ , \tag{5.5.3}$$

with $Z_F^{\mathrm{I}}(\theta^{\mathrm{I}}) = \int \exp(\langle s^{\mathrm{I}}(x^{\mathrm{I}})|\theta^{\mathrm{I}}\rangle) d\lambda^{\mathrm{I}}(x^{\mathrm{I}})$. The generic conjugate priors with hyperparameters $\lambda \in \mathbb{R}_+$, $y^{\mathrm{I}} \in \Omega_y^{\mathrm{I}}$ are given by

$$G^{\mathrm{I}}(\theta^{\mathrm{I}}|\lambda, y^{\mathrm{I}}) = \frac{1}{Z_G^{\mathrm{I}}(\lambda, y^{\mathrm{I}})} \exp(\langle \theta^{\mathrm{I}}|y^{\mathrm{I}}\rangle - \lambda \log(Z_F^{\mathrm{I}}(\theta^{\mathrm{I}}))) \ , \tag{5.5.4}$$

with partition function $Z_G^{\mathrm{I}}(\lambda, y^{\mathrm{I}}) = \int \exp(\langle \theta^{\mathrm{I}} | y^{\mathrm{I}} \rangle - \lambda \log(Z_F^{\mathrm{I}}(\theta^{\mathrm{I}})))d\theta^{\mathrm{I}}$. The hyperparameter space is $\Omega_y^{\mathrm{I}} = \mathbb{R}_+ \times (\Omega_y)^{\mathrm{I}}$. The posterior index, i. e. the mapping from prior to posterior parameters given the samples, is given according to (2.3.20) by

$$\big((\lambda, y^{\mathrm{I}}), x_1^{\mathrm{I}}, \dots, x_n^{\mathrm{I}}\big) \mapsto \big(\lambda + n, y^{\mathrm{I}} + \sum_{i=1}^{n} s^{\mathrm{I}}(x_i^{\mathrm{I}})\big) \ . \tag{5.5.5}$$

Whether the prior family is projective depends on the choice of the sufficient statistic, which enters the definition of the priors through Z_F. A result of some generality can be achieved by assuming decomposability of the sufficient statistic over subspaces. That is, for any $I \subset J$,

$$\langle s^{\mathrm{J}}(x^{\mathrm{J}}) | \theta^{\mathrm{J}} \rangle_{\Omega_\theta^{\mathrm{J}}} = \langle s^{\mathrm{I}}(x^{\mathrm{I}}) | \theta^{\mathrm{I}} \rangle_{\Omega_\theta^{\mathrm{I}}} + \langle s^{\mathrm{J}\setminus\mathrm{I}}(x^{\mathrm{J}\setminus\mathrm{I}}) | \theta^{\mathrm{J}\setminus\mathrm{I}} \rangle_{\Omega_\theta^{\mathrm{J}\setminus\mathrm{I}}} \ . \tag{5.5.6}$$

If this is the case, the partition function Z_F^{I} is factorial over component spaces, $Z_F^{\mathrm{J}}(\theta^{\mathrm{J}}) = Z_F^{\mathrm{I}}(\mathrm{P}_{\mathrm{J,I}}\theta^{\mathrm{J}}) Z_F^{\mathrm{J}\setminus\mathrm{I}}(\mathrm{P}_{\mathrm{J}\setminus\mathrm{I}}\theta^{\mathrm{J}})$. Consequently, both F^{I} and G^{I} are factorial over subspaces. The measures μ_Θ^{I} and $\mu^{\mathrm{I}}(\,.\,|\Theta^{\mathrm{I}} = \theta^{\mathrm{I}})$ are trivially projective, since e. g. $\mu^{\mathrm{J}}(\mathrm{R}_{\mathrm{J,I}} X^{\mathrm{I}} | \Theta^{\mathrm{J}} = \theta^{\mathrm{J}})$ is given by

$$\int_{\mathrm{R}_{\mathrm{J,I}} X^{\mathrm{I}}} F(x^{\mathrm{J}} | \theta^{\mathrm{J}}) dx^{\mathrm{J}} = \int_{X^{\mathrm{I}}} F(x^{\mathrm{I}} | \theta^{\mathrm{I}}) dx^{\mathrm{I}} \int_{\Omega^{\mathrm{J}\setminus\mathrm{I}}} F(x^{\mathrm{J}\setminus\mathrm{I}} | \theta^{\mathrm{J}\setminus\mathrm{I}}) dx^{\mathrm{J}\setminus\mathrm{I}} \ . \tag{5.5.7}$$

The factorial structure results in independent increment process in the projective limit. A simple example of a sufficient statistic that violates (5.5.6) is the quadratic statistic $s(x) = (x, xx')$ of the Gaussian. The only Gaussian processes representable in this form are Brownian motions, with or without drift.

If E is chosen countable, then by Prop. 36, an infinite-dimensional regular Bayesian model is uniquely defined on the limit space Ω_x^{E}, and conjugate by Cor. 39. Any measurable mapping $s^{\mathrm{E}} : \Omega^{\mathrm{E}} \to \Omega_\theta^{\mathrm{E}}$ satisfying (5.4.13) constitutes a sufficient statistic for the posterior of Θ^{E}, and the posterior under observations $x_1^{\mathrm{E}}, \dots, x_n^{\mathrm{E}}$ is

$$\mu^{\mathrm{E}}\big(\Theta^{\mathrm{E}} | Y^{\mathrm{E}} = (\lambda + n, y + \sum_i s^{\mathrm{E}}(x_i^{\mathrm{E}}))\big) \ . \tag{5.5.8}$$

5.5.3 Nonparametric Bayesian Mallows Model

[1] Meilă and Bao (2008) have suggested the extension of a Mallows rank model to a distribution on infinite permutations by means of a limit argument. Their model is essentially equivalent to the likelihood component $\mu_{X|\Theta}$ in the example presented here.

The previous examples described well-known standard models. To consider a non-standard model constructed by extension of an exponential family, we again take up the Fligner-Verducci model for permutations, which was used in Sec. 3.1 to cluster rank data. For fixed dimension (number of items) r, permutations in the symmetric group $\mathbb{S}(r)$ represent preference rankings of either exactly r objects, or for partial observations, a subset out of r objects. A nonparametric Bayesian model on infinite rankings is capable of representing the task to rank a person's favorite objects out of an unspecified number, finite or infinite. The way in which the Fligner-Verducci distribution models partial observations (cf. Sec. 3.1) ties in beautifully with the nonparametric Bayesian idea of adaptive model complexity.[2]

In particular, this example is intended to illustrate an alternative approach to the use of a sufficient statistic s, which is used here as a "preprocessing step" that maps observations into Ω_x. The random variable X is regarded as the image of a random observation under s. The approach can significantly simplify the formulation of the problem when the actual observation domain does not fit the structure of a metrizable space. Consider what it would take to rigorously formulate a projective limit construction directly on permutation-valued random variables. The group does not fit the extensions theorem's condition of a Polish topological space. We would have to identify a suitable surrogate topology and rederive the extension theorem for the group case. Moreover, for any non-standard type of random variable we may consider, we would have prove another version of the extension theorem. By choosing a suitable sufficient statistic that maps into a Polish space, the problem is reduced to existing results.

The Sufficient Statistic and its Properties

Again consider an infinite index set E, which will here be assumed to be countable and totally ordered. E^* denotes the set of finite subsets of E. Intuitively, the elements of E index items, and each $I \in E^*$ corresponds to a finite subset of items. We shall write $\mathbb{S}^{\mathrm{I}}$ for the permutation group of the set I. For all I of size $r \in \mathbb{N}$, the groups $\mathbb{S}^{\mathrm{I}}$ are isomorphic to the standard symmetric group $\mathbb{S}(r)$, but we have to distinguish between different items if $I_1 \neq I_2$, even if the sets are of the same size. In particular, $\mathbb{S}(r) = \mathbb{S}^{\mathrm{I}}$ for

[2] Incidentally, there is a much deeper connection between models used in Bayesian nonparametrics and the infinite symmetric group, apart from the construction described here and the obvious connection by infinite exchangeability: The representation theory of the infinite symmetric group has been studied thoroughly in a series of works by S. V. Kerov and A. M. Vershik. A class of measures arising naturally from the characters of the representations considered by Olshanski (2003) and Kerov *et al.* (2004) turn out to Poisson-Dirichlet processes.

$I = \{1, \dots, r\}$. For any permutation $\pi \in \mathbb{S}_r$, define a statistic

$$s_j(\pi) := \sum_{l=j+1}^{r} \mathbb{I}\{\pi^{-1}(j) > \pi^{-1}(l)\} \,, \tag{5.5.9}$$

where π^{-1} denotes the inverse of π in $\mathbb{S}(r)$ and $\mathbb{I}$ the indicator function of a set. When the statistics above are evaluated on an element $\pi \in \mathbb{S}(r)$, the r components can be collected in a vector, which will be denoted $s^r(\pi) := (s_1(\pi), \dots, s_r(\pi))$. (The last component is always trivial, but its inclusion simplifies notation.) Then the mapping $s^r : \mathbb{S}_r \to s^r(\mathbb{S}(r)) \subset \mathbb{R}^r$ is one-to-one (Fligner and Verducci, 1986). For $I \in E^*$, let s^{I} be the function induced by s^r. which takes $\mathbb{S}^{\mathrm{I}}$ into $\mathbb{R}^{\mathrm{I}}$. The function is evaluated by mapping $\mathbb{S}^{\mathrm{I}}$ isomorphically to $\mathbb{S}(r)$, and evaluating (5.5.9). More explicitly, consider a finite subset I. The order relation on E induces a unique reindexing of the elements, by assigning indices $1, \dots, |I|$ such that the order of elements is preserved (e. g. for $I = \{i_1, i_2, i_5, i_7\}$, relabel the last two elements as $i_5 \to i_3$ and $i_7 \to i_4$). The same is done for the rank positions. Each ranking $\pi^{\mathrm{I}} \in \mathbb{S}^{\mathrm{I}}$ then translates into a uniquely defined element of $\pi^r \in \mathbb{S}(r)$. If the subspace $\mathbb{R}^{\mathrm{I}}$ is identified with $\mathbb{R}^r$, then $s^{\mathrm{I}}(\pi^{\mathrm{I}}) = s^r(\pi^r)$.

Now consider the effect of projection. Let $I, J \in E^*$ with $I \subset J$. Choose a permutation $\pi^{\mathrm{J}} \in \mathbb{S}^{\mathrm{J}}$, and let $x^{\mathrm{J}} := s^{\mathrm{J}}(\pi^{\mathrm{J}})$. The projection $x^{\mathrm{I}} = \mathrm{P}_{\mathrm{J,I}} x^{\mathrm{J}}$ has two useful properties. First, let $\tilde{x}^{\mathrm{J}} = s^{\mathrm{J}}(\tilde{\pi}^{\mathrm{J}})$ be another element of $s^{\mathrm{J}}(\mathbb{S}^{\mathrm{J}})$. Then its projection coincides with that of x^{J} if and only if $\tilde{\pi}^J$ is identical to π^{J} on all positions $j \in I$. If the restriction of π^{J} to the positions indexed by the subset I is denoted $\pi^{\mathrm{J}}|_{\mathrm{I}}$, the set of such $\tilde{\pi}^{\mathrm{J}}$ is just the right coset of $\pi^{\mathrm{J}}|_{\mathrm{I}}$ in $\mathbb{S}^{\mathrm{J}}$, denoted $C^{\mathrm{J}}(\pi^{\mathrm{J}}|_{\mathrm{I}})$. The set is the analogue of the consistent set $C(\pi)$ discussed in Sec. 3.1.2. Then for any $\tilde{x}^{\mathrm{J}} = s^{\mathrm{J}}(\tilde{\pi}^{\mathrm{J}})$, we have $\mathrm{P}_{\mathrm{J,I}} \tilde{x}^{\mathrm{J}} = x^{\mathrm{I}}$ if and only if $\tilde{x}^{\mathrm{J}} \in C^{\mathrm{J}}(\pi^{\mathrm{J}}|_{\mathrm{I}})$. Assume that X^{I} is a subset of the embedding space $\mathbb{R}^{\mathrm{I}}$ that contains a single image $x^{\mathrm{I}} = s^{\mathrm{I}}(\pi^{\mathrm{I}})$ of a permutation. Then $s^{\mathrm{J}}(\mathbb{S}^{\mathrm{J}}) \cap \mathrm{R}_{\mathrm{J,I}} X^{\mathrm{I}} = s^{\mathrm{J}}(C^{\mathrm{J}}(\pi^{\mathrm{I}}))$. That is, the preimage of X^{I} under projection in the metric embedding space contains the images under s^{J} of those and only those permutations of J that are consistent with π^{I}.

The sufficient statistic mapping of the groups $\mathbb{S}^{\mathrm{I}}$ into the spaces $\mathbb{R}^{\mathrm{I}}$ thus results in a consistent projective structure, a fact perhaps not quite obvious initially for the embedding into metric product space of a decidedly non-geometric structure (a non-commutative finite group). Intuitively, this is due to the fact that the embedding aligns preimages under projection with right cosets of partial permutations. For $I \subset J$ fixed, the right cosets of all restrictions $\pi^{\mathrm{J}}|_{\mathrm{I}}$ form a partition of the group $\mathbb{S}^{\mathrm{J}}$ (they are pairwise disjoint and their union is the whole group). This behavior is consistent with preimages under projection in metric product space: The preimages $\mathrm{R}_{\mathrm{J,I}} x^{\mathrm{I}}$

of points in $\mathbb{R}^{\mathrm{I}}$ are pairwise disjoint for different points, and their union is just the space $\mathbb{R}^{\mathrm{J}}$. Consistency is ensured by the properties of s^{J}, which is defined such that (i) it is one-to-one (disjoint sets do not get mixed) and (ii) different positions in the permutations correspond to orthogonal axes in product space.

Finite-Dimensional Bayesian Model

For a suitably chosen parameter space $\Omega_\theta^{\mathrm{I}}$, the sufficient statistic s^{I} induces a conjugate Bayesian model, by choosing the canonical exponential family model for s^{I} and the induced conjugate prior. The parameter spaces considered here will be the sets $\Omega_\theta^{\mathrm{I}}(\epsilon)$, defined for any given $\epsilon > 0$ as $\Omega_\theta^{\mathrm{I}}(\epsilon) := \{\theta^{\mathrm{I}} \in \mathbb{R}^{\mathrm{I}} | \forall j \in I : \theta_j > \epsilon\}$. This is just the positive orthant, constantly bounded away from all axes, and hence open and convex. By choosing the positive orthant, the resulting exponential family model is the Fligner-Verducci distribution (Fligner and Verducci, 1986). The likelihood component is the exponential family distribution

$$F^{\mathrm{I}}(\pi^{\mathrm{I}}|\sigma^{\mathrm{I}},\theta^{\mathrm{I}}) := \frac{1}{Z^{\mathrm{I}}(\theta)} \exp\Big(\big\langle s^{\mathrm{I}}(\pi^{\mathrm{I}}(\sigma^{\mathrm{I}})^{-1})|\theta^{\mathrm{I}}\big\rangle_{\mathbb{R}^{r-1}}\Big) , \tag{5.5.10}$$

with partition function $Z^{\mathrm{I}}(\theta^{\mathrm{I}}) = \prod_{j=1}^{r-1}(1 + (r-j)\exp(-\theta_j^{\mathrm{I}}))$. For a discussion of the model's properties and a derivation of the partition function, cf Fligner and Verducci (1986). It will be convenient to rewrite the probability mass function above as a density w. r. t. Lebesgue measure on $\mathbb{R}^{\mathrm{I}}$,

$$\tilde{F}(x^{\mathrm{I}}|\theta^{\mathrm{I}}) = \frac{1}{\tilde{Z}_F} \exp\left(-\left\langle x^{\mathrm{I}}|\theta^{\mathrm{I}}\right\rangle\right) \delta_{s^{\mathrm{I}}(\mathbb{S}^{\mathrm{I}})}(x^{\mathrm{I}}) . \tag{5.5.11}$$

The generic conjugate prior on t with hyperparameters $\lambda \in \mathbb{R}_+$, $\mathbf{y} \in \mathbb{R}^{r-1}$ is given by the densities

$$G^{\mathrm{I}}(\theta^{\mathrm{I}}|\lambda,\mathbf{y}^{\mathrm{I}}) \propto \exp\Big(\langle\theta^{\mathrm{I}}|\mathbf{y}^{\mathrm{I}}\rangle_{\mathbb{R}^{r-1}} - \lambda \log Z^{\mathrm{I}}(\theta^{\mathrm{I}})\Big) . \tag{5.5.12}$$

Formally, the model components as defined in Sec. 5.3 are $\Omega_x = \Omega_\theta = \Omega_y = \mathbb{R}$, with Borel algebras $\mathcal{B}_x$, $\mathcal{B}_\theta$ and $\mathcal{B}_y$, and the prior measures μ_Θ^{I} have densities G^{I} w. r. t. Lebesgue measure on $\Omega_\theta^{\mathrm{I}}$. Since the model is a canonical exponential family model, the hyperparameter space for given $I \in E^*$ is $\mathbb{R} \times \mathbb{R}^{\mathrm{I}}$, where the first axis accommodates the prior dispersion λ. The posterior index s^{I} under observations $\pi_1^{\mathrm{I}}, \dots, \pi_n^{\mathrm{I}}$ is generated pointwise by the mapping

$$((\lambda,\mathbf{y}), s^{\mathrm{I}}(\pi_1^{\mathrm{I}}), \dots, s^{\mathrm{I}}(\pi_n^{\mathrm{I}})) \mapsto (\lambda + n, \mathbf{y} + \sum_{i=1}^{n} s^{\mathrm{I}}(\pi_i^{\mathrm{I}})) . \tag{5.5.13}$$

Infinite-Dimensional Extension

As the discussion of the sufficient statistic above shows, s^{I} satisfies (5.5.6), and therefore is projective. The independent increment property of resulting limit process corresponds to a well-known mutual independence property of the component statistics s^{I}_j (Fligner and Verducci, 1986). The infinite extension is constructed on the infinite replication of the embedding space $\Omega_x = \mathbb{R}$. Since the lower bound ϵ of the parameter vectors is chosen uniformly over all $I \in E^*$ and constitutes a point-wise property, it carries over to the infinite limit. Since the model on Ω^{I}_x corresponds to permutations of the elements of I, the infinite extension model defines a Bayesian model for permutations of the infinite index set E, that is, on the set $\bar{\mathbb{S}}(E)$ of all bijections of the set E. The parameters estimated by the Bayesian model are weight functions $\theta : E \to \mathbb{R}_{>\epsilon}$. The nonparametric Bayesian character of the model is its way to cope with partial observations, that is, only a finite number of positions of the permutation is observed (such as a partial ranking of an infinite number of objects). In this case, estimates for the remaining positions are filled in by prior assumption.

5.6 Discussion

The questions considered here are motivated by interest in machine learning, rather than statistics. Bayesian nonparametric statistics has concentrated its efforts on the definition of measures on the set of probability measures, motivated in particular by the search for universal priors. In contrast, the definition of models on different domains has attracted interest in machine learning and cognitive science.

Modeling

The picture that emerges shows that despite the subtleties of infinite-dimensional spaces, and despite the many well-studied properties of stochastic processes which set them apart from ordinary multivariate distributions, a number of fundamental model properties carry over from the parametric to the nonparametric case. Models that are analytically tractable in a conjugate sense arise only from marginal Bayesian systems which are themselves conjugate, and hence essentially from exponential families. If there is a statistic s^{E} in the infinite-dimensional case that extends the finite-dimensional sufficient statistics in the sense of (5.4.13), then it is sufficient for the extended posterior. The structure of the limit posterior is then

Marginals (d-dim)	**Projective limit**	**Observations (limit)**
Bernoulli/Beta	IBP/Beta process	Binary arrays
Multin./Dirichlet	CRP/DP	Discrete distributions
Gaussian/Gaussian	GP/GP	(continuous) functions
Mallows/conjugate	Example Sec. 5.5.3	Bijections $\mathbb{N} \to \mathbb{N}$

analogous to parametric posteriors in conjugate exponential families,

$$\mu_\Theta(\Theta^{\mathrm{E}}|X_1 = x_1, \ldots, X_n = x_n) = \mu_\Theta\Big(\Theta^{\mathrm{E}}\Big|Y^{\mathrm{E}} = \big(\lambda + n, y + \sum_{i=1}^{n} s^{\mathrm{E}}(x_i^{\mathrm{E}})\big)\Big) \tag{5.6.1}$$

The construction of such models is, to a certain degree, generic. To construct a distribution on an infinite-dimensional object, start by consider its finite-dimensional restrictions or counterparts. That may be a vector (for functions), a matrix (for linear operators), a finite graph (for an infinite one) etc. Then, proceed as follows:

1. Choose either an available multivariate exponential family model, or define one by choosing a set of sufficient statistics which measure whatever properties are of interest.
2. Choose a natural conjugate prior.
3. Consider the resulting Bayesian systems on arbitrary finite dimensions. Check that these systems are projective. It is sufficient to check that the prior and either the sampling distribution or the posterior distribution are projective.
4. If the finite-dimensional systems are projective, their sufficient statistics will have the same functional form on arbitrary dimensions. Define a mapping s^{E} of analogous functional form on the infinite-dimensional sample space, and verify that it extends the sufficient statistics (i. e. that it satisfies either (5.4.13) or (5.4.18)).

Then there is a infinite-dimensional Bayesian limit system, defined uniquely (up to equivalence). If the mapping s^{E} extending the sufficient statistics can be identified, it is Bayesian sufficient for the limit system, and specifies the posterior when prior and data are given.

Model Taxonomy

The results above map out a taxonomy of Bayesian nonparametric models that closely resembles the familiar landscape of models in the parametric case. Three types of models are distinguished:

Type I: Extensions of systems with conjugate priors.
This is the case discussed above in much detail. Some examples are given in Tab. 5.6.

Type II: Extensions of systems with mixtures of conjugate priors.
If the prior is a finite mixture of conjugate priors, the posterior is the corresponding mixture (with identical weights) of the individual conjugate posteriors. In this case, the results above can be applied individually to each mixture component. Mixtures of conjugate priors in the finite-dimensional parametric case have been studied by Dalal and Hall (1983). Continuous mixtures will probably require further study. Results should remain applicable if the dimension of the mixing variable (the variable Z in the notation of Ch. 2) is independent of the dimension of observations and model parameters. For example, if the covariance matrices Σ of a projective family of Gaussians are scaled by some $\tau \in \mathbb{R}_+$, then τ can be used as a mixing variable. This variable is independent of the dimension of the Gaussian model. Mixing any of the resulting finite-dimensional Gaussian models against a gamma distribution on τ produces a Student t-model (cf Example 26). If the same gamma distribution is used on all models of the family, the extension model is a continuous mixture of Gaussian processes, one for each value of τ, against the same gamma distribution. Such a mixture of Gaussian processes has been introduced as a *Student t-process* by O'Hagan (1991).

Type III: Non-conjugate priors.
In the finite-dimensional case, non-conjugate posteriors usually require either analytic approximations or numerical methods. One (speculative) approach to non-conjugate posteriors in nonparametric Bayesian models may be the extension of approximate posteriors. That is, to define a family of Bayesian systems on finite-dimensional subspaces, each with approximate posteriors, such that both the priors and the posterior approximations are projective. Then the system can in principle be extended in much the same way as a conjugate one. In particular, if a closed-form mapping from the data to the parameters of the approximating posterior distribution is available, this could be regarded as a sufficient statistic. However, it should be pointed out that, in contrast to sufficiency and conjugacy, approximation quality of distributions is *not* a property that can be expected to carry over to the infinite-dimensional case without further provisions.

Caveats

Not all important model properties are directly extendable. Not all important properties of the finite-dimensional models are preserved under extension, even though conjugacy and sufficiency are. This concerns in particular two types of properties: Analytic properties of random functions drawn from the limit model, such as smoothness, and convergence properties of the probability models, including consistency. Posterior consistency is a particular issue: The Dirichlet process can produce inconsistent estimates, regardless of the fact that its finite-dimensional marginals are consistent (Diaconis and Freedman, 1986). The properties studied above do not directly relate to the problem of consistency in Bayesian nonparametric models – consistency is a property of the posterior defined by the model, whereas sufficiency and conjugacy concern the computation of this posterior, regardless of its properties. Analytic properties concern cases in which the infinite-dimensional object can be interpreted as a function on a continuous domain (as in the Gaussian process case). Construction by means of the extension theorem defines the random function in a point-wise manner (each marginal specifies the function's restriction to finite subset of points). Analytic properties such as continuity or smoothness, which involve the notion of an open neighborhood, are not expressible in this manner.

Uncountable index sets. If the index set E is uncountable, the extended measure lives on a σ-algebra $\mathcal{B}^{\mathrm{E}}$ that does not contain the singletons (the one-point sets). This poses a principal problem for Bayesian methods, because it means that application of the measure to a sample observation (which is a point) is not defined. Constructions of systems on random quantities over e. g. the real line are thus of limited use, which seems somewhat alarming, considering that both the Gaussian and the Dirichlet process are usually considered for random functions or measures on the real line. However, neither model has a genuinely uncountable number of degrees of freedom, in the following sense: The Gaussian process is considered only for continuous functions. A continuous function on $\mathbb{R}$ is completely determined by its values on a dense countable subset. A Gaussian process model on continuous functions over $\mathbb{R}$ can therefore be constructed by extending a Gaussian model with $E = \mathbb{Q}$ as the index set. The argument remains valid for any domain with a dense subset that is countable (and hence for any separable metric space). Similarly, as Ghosh and Ramamoorthi (2002) point out, the Dirichlet process on $\mathbb{R}$ can be constructed on $E = \mathbb{Q}$, because the cumulative distribution functions on $\mathbb{R}$ admit a bimeasurable isomorphism to the set of their restrictions on $\mathbb{Q}$. (That is, two cumulative distribution functions with domain $\mathbb{R}$ define the same probability measure if and only if their restrictions to $\mathbb{Q}$ are identical.) The examples suggest that many problems of practical interest admit a surrogate construction of countable dimension. Since repetitive observation is an inherently countable process, it is indeed hard to imagine how a model with an uncountable effective number of degrees of freedom should admit a meaningful notion of asymptotic behavior. As a general rule of thumb, constructions involving uncountable *domains* are easily handled by standard measure theory. Constructions involving uncountably repeated *operations*, such as the uncountable product of axes defined by $E = \mathbb{R}$, are typically problematic, since the definitions of measures and σ-algebras admit only countably infinite operations.

Appendix A

Reference Results

The following are results used in proofs. They are reproduced here for reference, without proofs or further explanation.

Existence of Regular Conditional Probabilities

Conditional probabilities (cf App. B) are not generally guaranteed to have a regular version, i. e. a version that is a probability measure almost everywhere. But there is always a regular version if the topology of the space is not much more complicated than that of the real line, as the following theorem ensures.

Theorem 44. *Let $X : (\Lambda, \mathcal{A}) \to (\Omega, \mathcal{B}(\Omega))$ such that Ω is Polish. Then for every σ-subalgebra $\mathcal{C} \subset \mathcal{A}$, there is a regular conditional distribution of X given $\mathcal{C}$. Moreover, there is an $\mathbb{P}$-null set N such that any two such regular condtional distributions coincide for every $\omega \in \complement N$.*

The standard proof of this theorem (see for example Bauer, 1996) establishes a number of intermediate results, some of which are used in the proof of lemma 34, and summarized here for reference.

Lemma 45. *Let $X : (\Lambda, \mathcal{A}) \to (\Omega, \mathcal{B}(\Omega))$ be a random variable, with $(\Omega, \mathcal{B}(\Omega))$ a Borel space. Then for every σ-subalgebra $\mathcal{C} \subset \mathcal{A}$, the following holds:*

1. *There exists a countable algebra $\mathcal{G}$ that generates $\mathcal{B}(\Omega)$.*
2. *Define a function P by choosing a version of $\mathbb{E}\left[\mathbb{I}_A | \mathcal{C}\right]$ for every $A \in \mathcal{G}$, and setting $P(A, \omega) := \mathbb{E}\left[\mathbb{I}_A | \mathcal{C}\right](\omega)$. Then there is an $\mathbb{P}$-null set N such that $A \mapsto P(A, \omega)$ is a measure on the algebra $\mathcal{G}$ for all $\omega \in \complement N$.*

3. *The unique extension of $P(\,.\,,\omega)$ from $\mathcal{G}$ to $\mathcal{C}^E$ (as given by the extension theorem for measures*[1]*) is measurable w. r. t. ω.*

Martingale Convergence

Those parts of Ch. 5 which consider the projective limit with respect to a condition are greatly simplified by the following observation: The σ-algebras, on which the measures are conditioned, form nested sequences over different dimensions. Increasing the dimension of subspaces corresponds to a sequence of σ-algebras of increasing resolution. Such successively refined sequences form a fundamental tool in the theory of martingales (where they are called *filtrations*), and standard results on martingale convergence are applicable in the proofs of Ch. 5. The two relevant theorems are given below. The actual convergence result is the second one, which requires the martingale involved to satisfy a condition known as *uniform integrability.* The first theorem is needed only to guarantee that this condition is satisfied for all integrable real-valued random variables, and hence in particular for the 0-1 variables $\mathbb{I}_A$ in conditional probabilities.

Theorem 46. *Let X be a real-valued, integrable random variable on $(\Lambda, \mathcal{A}, \mathbb{P})$. Let I be a partially ordered index set and $(\mathcal{C}_i)_{i \in I}$ a filtration in $\mathcal{A}$. Then*

$$(\mathbb{E}\left[X|\mathcal{C}_i\right], \mathcal{C}_i)_{i \in I} \tag{A.0.1}$$

is a uniformly integrable martingale.

Theorem 47. *Let I be an index set with its partial order $\leq$ chosen such that, for any $i_1, i_2 \in I$, there is some $i_3 \in I$ with both $i_1 \leq i_3$ and $i_2 \leq i_3$. Let $(X_i, \mathcal{C}_i)_{i \in I}$ be a uniformly integrable martingale. Then there exists one and (up to equivalence) only one integrable, $\mathcal{C}_\infty$-measurable random variable X_∞ such that*

$$\forall i \in I : \qquad X_i = \mathbb{E}\left[X_\infty|\mathcal{C}_i\right] \qquad a.s. \tag{A.0.2}$$

The condition on the index set in the second theorem restricts the arbitrariness of choice of the partial order relation $\leq$ on I, and is always satisfied if I is totally ordered, or if I is an upper semilattice. Its relevance for filtrations is that, due to the partial order, two σ-algebras in the filtration may not be comparable, that is, if their indices are not comparable by the partial order relation, neither system is guaranteed to be a subset of the other. For

[1] "Extension theorem" here refers to Caratheodory's theorem on the extension of an arbitrary measure from a generating algebra to a σ-algebra, rather than Kolmorogorov's theorem on extension to infinite dimenions.

the above theorem (and many other convergence results), such "branches" in the sequence are admissable, as long as there is another σ-algebra further down in the sequence which contains both.

The de Finetti Theorem

De Finetti's theorem has become generally well-known in the machine learning community. Nonetheless, the requirements of the theorem are interesting with respect to Kolmogorov's extension theorem and the discussions in Ch. 5, because they involve random variables with values in Polish or Borel spaces. The following version is due to Kallenberg (2005). An infinite sequence $X_1, X_2, \dots$ of random variables is called *contractable* if its distribution is invariant under restriction to an arbitrary infinite subsequence $X_{i_1}, X_{i_2}, \dots$ (where $i_1 < i_2 < \dots$). It is called *exchangeable* if its distribution is invariant under arbitrary permutations of finite subsets of indices.

Theorem 48. *Let $X = (X_1, X_2, \dots)$ be an infinite sequence of random variables with values in a measurable space $(\Omega, \mathcal{A})$. If the space is Borel, the following three conditions are equivalent:*

1. *X is contractable.*
2. *X is exchangeable.*
3. *X is conditionally i.i.d.*

If the space is not Borel, (1)$\Leftrightarrow$(2)$\Leftarrow$(3).

The equivalence between contractability and exchangeability does not hold for finite sequences. The consequence of implication (2)$\Rightarrow$(3) is that any infinite exchangeable sequence X can be represented as a mixture of product models: If $X = (X_1, X_2, \dots)$ is exchangeable, there is some non-trivial σ-subalgebra $\mathcal{C}$ such that the joint conditional distribution of each subsequence $X^n = (X_1, \dots, X_n)$ decomposes as

$$\mu_{X^n}(X^n|\mathcal{C}) = \prod_{i=1}^{n} \mu_{X_1}(X_i|\mathcal{C}) \; . \tag{A.0.3}$$

In the particular case $\Omega = \mathbb{R}$, the set $\mathcal{M}^1_+(\Omega)$ be the set of all probability measures on the sample space Ω can be identified with the set of distribution functions $\mathcal{F}(\mathbb{R})$. In this case, the theorem states that any infinite sequence is exchangeable (and contractable) if and only if the joint probability P_n of of each n-subsequence is

$$P_n(x_1, \dots, x_n) = \int_{\mathcal{F}(\mathbb{R})} \prod_{i=1}^{n} F(x_i) d\nu(F) \; . \tag{A.0.4}$$

Given P_n (for all n), ν is uniquely determined: If $\hat{F}_n$ denotes the empirical distribution function of observed values $x_1, \dots, x_n$, then for *any* sequence $(x_i)_{i \in \mathbb{N}}$,

$$\nu\big(\lim_{n\to\infty} \hat{F}_n\big) = \lim_{n\to\infty} P_n(x_1, \dots, x_n) \; . \tag{A.0.5}$$

The representation theorem carries over from discrete stochastic processes $(X_i)_{i \in \mathbb{N}}$ with exchangeable values to continuous stochastic processes with exchangeable increments. The term *increment* refers to the increment $(X_t - X_s)$ on any finite intervall $[s, t] \subset \mathbb{R}_+$. In many regards, the closest analogue of an i.i.d. property of a countable set of observations in continuous time is the stationary independent increment property of Lévy processes. The following theorem states that this property holds *conditionally* on some σ-algebra if the increments of the process are exchangeable.

Theorem 49 (Bühlmann, 1960)**.** *Let $(X_t)_{t \in \mathbb{R}_+}$ be a stochastic process for which $X_s \xrightarrow{P} X_t$ whenever $s \to t$. Then the increments of X are exchangeable (and even contractable) if and only if they are conditionally stationary and independent, given some σ-subalgebra $\mathcal{C}$.*

Processes for which $s \to t$ implies $X_s \xrightarrow{P} X_t$ are called *continuous in probability.* Theorem 49 also holds, with some modifications, if the index set is compact (e. g. $[0, 1]$ instead of $\mathbb{R}$), though the upper bound on the index range leads to considerable complications in the proof. The representation of X (which corresponds to the conditionally factorial representation above) then explicitly contains a Brownian bridge, and the result includes Donsker's theorem as a special case (Kallenberg, 1997).

Appendix B

Conditioning

Conditional probabilities are usually represented in machine learning in terms of conditional densities, with no need to consider the underlying theory in terms of measures. Conditioning probability measures on arbitrary events (in particular including those of measure zero) is slightly more subtle. The following is a brief review of conditional expectations and conditional distributions, which mostly serves to define notation. The concepts are elementary in probability theory. For a comprehensive introduction, see Bauer (1996) and Loève (1977b), or Kallenberg (1997) for a more abstract treatise.

Conditioning in terms of measure theory is a straightforward matter if the condition event is not a null set: Define $\mu(A|B) = \frac{\mu(A \cap B)}{\mu(B)}$. But conditioning in statistics likely as not means conditioning on a single value or observation, which typically has measure zero. Kolmogorov noticed that, if the conditionals of a measure are known on all non-null sets, then values of the measure conditional on the null sets can be deduced (at least almost everywhere) – they cannot be arbitrary, because that would change the behavior on the non-null sets as well. The underlying concept is that of a weak definition, or weak identity, which regards two functions as identical if they integrate in the same manner. That two functions have the same integral on a given set is not much of a constraint, but if they integrate identically on *all* sets, they can differ only on a set of measure zero.

Conditional Expectation

The expectation of a random variable, obtained by integrating it over all of Ω, can be refined by subdividing Ω into a partition, and taking expectations over the parts of this partition. The individual expecations can be joined in a single function by switching elements on and off by means of an indicator function. This can be regarded as a smoothing of the random variable: The coarser the resolution of the chosen partition, the stronger the smoothing effect. Maximal smoothing is attained for a partition consisting only of Ω, and results in the standard expectation. If the partition is a generator of the underlying σ-algebra, there is not smoothing, and the original random variable is recovered.

Countable case. Let $X : (\Omega_1, \mathcal{A}_1) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ be a random variable, μ a probability measure on $(\Omega_1, \mathcal{A}_1)$ and $\mathcal{D} := \{C_i\}_{i \in J}$ a countable partition of Ω. For any C_i with $\mu(C_i) > 0$, denote by $\mathbb{E}_{C_i}[X]$ the expectation of X over C_i:

$$\mathbb{E}_{C_i}[X] := \frac{1}{\mu(C_i)} \int_{C_i} X(\omega) d\mu(\omega) \tag{B.0.1}$$

That is, $\mathbb{E}_{C_i}[X]$ is the expectation of X calculated w. r. t. the probability measure $\frac{\mu(\,.\cap C_i)}{\mu(C_i)}$.

The first step towards conditional expectation is to define a random variable $X_{\mathcal{D}}$ on the non-null sets of $\mathcal{D}$: Let $J_0 \subset J$ be the indices of the μ-null sets in $\mathcal{D}$, and define

$$X_{\mathcal{D}}(w) := \sum_{i \in J \setminus J_0} \mathbb{E}_{C_i}[X] \mathbb{I}_{C_i} . \tag{B.0.2}$$

This is a random variable that takes value $\mathbb{E}_{C_i}[X]$ for $\omega \in C_i$, and is undertermined in case of the null event $\omega \in \bigcup_{i \in J_0} C_i$. Since J is countable, $X_{\mathcal{D}}$ is an elementary function or "step function", the right-hand side of (B.0.2) is just its canonical representation, and its integral is a sum over J. The mapping $X_{\mathcal{D}}$ is almost the conditional expectation of X given $\mathcal{D}$. What remains to be done is to generalize from the partition to a σ-algebra, and to take care of the null sets.

Considering the σ-algebra $\mathcal{C} := \sigma(\mathcal{D})$ generated by the partition, the expectation over any $C \in \mathcal{C}$ can be computed immediately from $X_{\mathcal{D}}$: Any $C \in \mathcal{C}$ is a combination of some of the C_i, i.e. there is a subset $I \subset J$ such that $C = \bigcup_{i \in I} C_i$. Hence,

$$\mathbb{E}_C[X] = \int_C X(\omega) d\mu(\omega) = \sum_{i \in I} \mathbb{E}_{C_i}[X] \mu(C_i) . \tag{B.0.3}$$

It seems justified to rename the r.v. $X_{\mathcal{D}}$ to $X_{\mathcal{C}}$, or better still $\mathbb{E}[X|\mathcal{C}]$. The conditional expectation is explained on all of $\mathcal{C}$, including the null sets, by allowing it to assume arbitrary values on the null sets. The rationale is that, when evaluated on a random event, a meaningful value is assumed with probability one.

Definition 50 (Conditional expecation: Constructive definition)**.** Let the partition $\mathcal{D} = \{C_i\}_{i\in J}$ of Ω_1 be countable, and $J_0 \subset J$ be the indices of its null sets. Then any function $\mathbb{E}[X|\mathcal{C}]$ which takes values

$$\mathbb{E}[X|\mathcal{C}](\omega) := \sum_{i\in J\setminus J_0} \mathbb{E}_{C_i}[X]\,\mathbb{I}_{C_i}(\omega) \tag{B.0.4}$$

on the non-null region, and arbitrary values on the null region $\bigcup_{i\in J_0} C_i$, is called a *version of the conditional expectation* of X given $\mathcal{C}$.

General (possibly uncountable) case. The constructive, explicit definition above admits no direct generalization to the uncountable case. In the countable case, null sets can be ignored, by merit of their countable union once again being a null set. In the uncountable case, for a finite measure, almost all sets are null sets, and there union is generally non-null. (A Borel algebra on $\mathbb{R}$, for example, contains singletons as smallest null sets, the union of which is the entire space.) Allowing arbitrary values of the function on these sets would hence result in an arbitrary contribution to the integral. Generalization to the uncountable case is made possible by substituting an implicit definition of weak type, which in the countable special case is equivalent to the constructive definition above.

Definition 51 (Conditional expectation: Implicit definition)**.** Let $(\Omega_1, \mathcal{A}_1, \mu)$ be a probability space, $\mathcal{C}$ any sub-σ-algebra of $\mathcal{A}_1$, and X an $\mathcal{A}_1$-measurable, integrable random variable. Let $\mu\big|_{\mathcal{C}}$ denote the restriction of μ to $\mathcal{C}$. Any $\mathcal{C}$-measurable function $\mathbb{E}[X|\mathcal{C}]$ satisfying

$$\forall C \in \mathcal{C}: \qquad \int_C \mathbb{E}[X|\mathcal{C}]d\mu\big|_{\mathcal{C}} = \int_C X d\mu \tag{B.0.5}$$

is called a *version of the conditional expectation* of X given $\mathcal{C}$.

Note that $\mu\big|_{\mathcal{C}}$ and μ can be used interchangeably on the left hand side, since $\mathbb{E}[X|\mathcal{C}]$ is a $\mathcal{C}$-measurable function.

If the partition is countable and $\mathbb{E}[X|\mathcal{C}]$ defined in the construtive manner (B.0.4), then it satisfies (B.0.5) as well. For $C \in \mathcal{C}$, let I again denote the indices of C in the partition, $C = \bigcup_{i\in I} C_i$.

$$\begin{aligned}\int_C \mathbb{E}[X|\mathcal{C}](\omega)d\mu(\omega) &= \int_C \sum_{i\in J\setminus J_0} \mathbb{E}_{C_i}[X]\,\mathbb{I}_{C_i}(\omega)d\mu(\omega) \\ &= \int_C \sum_{i\in I\setminus J_0} \mathbb{E}_{C_i}[X]\,d\mu(\omega) = \sum_{i\in I\setminus J_0} \mathbb{E}_{C_i}[X]\int_{C_i} d\mu(\omega) \\ &= \sum_{i\in I\setminus J_0} \mathbb{E}_{C_i}[X]\,\mu(C_i) = \int_C X(\omega)d\mu(\omega)\end{aligned}$$

Unlike a constructive definition, an implicit one requires a discussion of existence. The following theorem shows that the definition is safe.

Theorem 52 (Existence and uniqueness of general conditional expectations)**.** *Let $(\Omega_1, \mathcal{A}_1, \mu)$ be a probability space. Then for any sub-σ-algebra $\mathcal{C}$ of $\mathcal{A}_1$, and $\mathcal{A}_1$-measurable, integrable function X, the conditional expectation $\mathbb{E}[X|\mathcal{C}]$ defined by (B.0.5) exists, and is unique modulo μ.*

Conditioning on a random variable. The conditional expectation given a $\mathcal{A}_1 - \mathcal{A}_2$-measurable function $Y : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ (i. e. a random variable) is defined as the conditional expectation given the σ-algebra induced by Y.

Definition 53 (Induced σ-algebra)**.** For any mapping T with values in $(\Omega, \mathcal{A})$, the the *σ-algebra induced by T*, denoted $\sigma(T)$, is the smallest σ-algebra which makes T $\sigma(T) - \mathcal{A}$-measurable. For a family of mappings $\{T_i\}_{i\in I}$ with values in $(\Omega_i, \mathcal{A}_i)$, the induced σ-algebra $\sigma(T_i, i \in I)$ is the smallest σ-algebra such that *each* mapping T_i is $\sigma(T_i, i \in I) - \mathcal{A}_i$-measurable.

For a single mapping, by the properties of the preimage operation, $T^{-1}(\mathcal{A})$ is a σ-algebra, and hence $\sigma(T) = T^{-1}(\mathcal{A})$. When considering a family of mappings, one has to consider the union $\bigcup_{i\in I} T_i^{-1}(\mathcal{A}_i)$ of all preimages, which is not generally a σ-algebra. Hence, the union is used as a generator, and $\sigma(T_i, i \in I) := \sigma(\bigcup_{i\in I} T_i^{-1}(\mathcal{A}_i))$.

Definition 54 (Conditional expectation given a random variable)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, and X and Y be random variables with values in $(\Omega_x, \mathcal{A}_x)$ and $(\Omega_y, \mathcal{A}_y)$, respectively. The conditional expectation of X given Y is

$$\mathbb{E}\left[X|Y\right] := \mathbb{E}[X|\sigma(Y)] \; . \tag{B.0.6}$$

For a family $\{Y_i | i \in I\}$ of random variables, with values in $(\Omega_i, \mathcal{A}_i)$, the conditional expectation of X given $\{Y_i | i \in I\}$ is

$$\mathbb{E}\left[X|Y_i, i \in I\right] := \mathbb{E}[X|\sigma(Y_i, i \in I)] \; . \tag{B.0.7}$$

Explanation. In general, a conditional expectation $\mathbb{E}[X|\mathcal{C}]$ resolves local (i. e. subset-conditional) expectations on all sets of the σ-algebra $\mathcal{C}$. Assume that an event $Y(\omega) \in A \in \mathcal{A}_y$ is observed. Conditioning on this event in $(\Omega_y, \mathcal{A}_y)$ requires pulling it back to $(\Omega, \mathcal{A})$, and hence conditioning on the event $Y^{-1}(A)$. The σ-algebra generated by the pulled-back events is just $\sigma(Y)$.

The conditional expecation $\mathbb{E}[X|Y]$ is a function of ω, which can be interpreted as follows: With some ado, we can derive a function $\mathbb{E}\left[X|Y = y\right]$, which maps y to a number. It can than be shown that, for any version $\mathbb{E}[X|Y]$ of the conditional expectation,

$$\mathbb{E}[X|Y](\omega) = \mathbb{E}\left[X|Y = Y(\omega)\right] \qquad \text{almost surely.} \tag{B.0.8}$$

Hence, the value of $\mathbb{E}[X|Y](\omega)$ can be understood as the expectation of X, given that Y assumes the value $Y(\omega)$.

Conditional Probability

Definition 55 (Conditional probability)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, $\mathcal{C}$ a sub-σ-algebra of $\mathcal{A}$. The *conditional probability* of an event $A \in \mathcal{A}$ given the hypothesis $\mathcal{C}$ is the $\mathcal{C}$-conditional expectation of the random variable $\mathbb{I}_A$,

$$\mu(A|\mathcal{C}) := \mathbb{E}[\mathbb{I}_A|\mathcal{C}]\ . \tag{B.0.9}$$

Explanation. The conditional expectation point of view is to define quantities from an integration perspective. Conditional expectation is defined as a random variable that integrates like a conditional expectation. Similarly, the conditional probability definition above would be justified by arguing that it integrates as a conditional probability should. The standard definition of conditional probability, $\mu(A|C) = \frac{\mu(A\cap C)}{\mu(C)}$, integrates as

$$\int_C \frac{\mu(A\cap C)}{\mu(C)} d\mu(\omega) = \mu(A \cap C)\ . \tag{B.0.10}$$

The same is true for the conditional probability $\mu(A|\mathcal{C})$ defined above, since

$$\begin{aligned}\int_C \mu(A|\mathcal{C}) d\mu(\omega) = \int_C \mathbb{E}[\mathbb{I}_A|\mathcal{C}] d\mu(\omega) = \int_C \mathbb{E}[\mathbb{I}_A|\mathcal{C}] d\mu|_{\mathcal{C}}(\omega) = \int_C \mathbb{I}_A(\omega) d\mu(\omega) \\ = \mu(A\cap C)\ .\end{aligned} \tag{B.0.11}$$

Conditional probability as a function. The integral perspective on conditional probabilities tents to distract from how they work when regarded as a function of ω, i.e. $\mu(A|\mathcal{C})(\omega)$. This is most easily illustrated by the elementary case, i.e. for a $\mathcal{C}$ generated by a countable partition $\{C_i\}$ consisting only of non-null sets. Fix A, and assume that ω assumes a value in the element C_0 of the partition. Then

$$\mu(A|\mathcal{C})(\omega) = \mathbb{E}_{C_i}\left[\mathbb{I}_A\right] = \frac{1}{\mu(C_0)} \int_{C_0} \mathbb{I}_A(\omega') d\mu(\omega') = \frac{\mu(A\cap C_0)}{\mu(C_0)}\ . \tag{B.0.12}$$

In the general case, we have to revert to the integral definition and understand $\mu(A|\mathcal{C})(\omega)$ as an object not intended to be evaluated at an isolated point, but to be integrated over a set.

Regular Conditional Distributions

Working with conditional distributions is complicated by the fact that a conditional distribution as defined in (B.0.9) is not guaranteed to be a probability measure almost everywhere. That is, the function $A \mapsto \mu(A|\mathcal{C})(\omega)$ need not be a probability measure for each ω. Conditional distributions which *are* probability measures for each ω are called *regular conditional probabilities.* For $\mathcal{C}$ fixed, a regular condional probability $\mu(A|\mathcal{C})(\omega)$ can be regarded as a function $K(\omega, A)$ of two arguments. Because K is a conditional expectation as in (B.0.9), it is always $\mathcal{A}$-measurable for fixed A, and because it is regular, it is a probability measure w. r. t. to A for fixed ω. But these are just the two properties that define a Markov kernel in functional analysis, and therefore, regular conditional probabilities can be represented by Markov kernels.

Definition 56 (Markov kernel)**.** Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be measurable spaces. A *Markov kernel* is a mapping

$$K : \Omega_1 \times \mathcal{A}_2 \to [0, +\infty] \qquad \text{(B.0.13)}$$

such that:

1. $\forall A \in \mathcal{A}_2 : \quad K(\,.\,, A)$ is $\mathcal{A}_1$-measurable.
2. $\forall \omega \in \Omega_1 : \quad K(\omega,\,.\,)$ is a *probability* measure on $\mathcal{A}_2$.

In general, a kernel defines a linear operator which maps functions to functions. A Markov kernel is a kernel which is normalized to define an operator which maps probability measures to probability measures. Image measures are kernels: If T is a measurable map, the image measure $(T(\mu))(A) = \mu(T^{-1}(A))$ is a kernel (which is independent of ω). Hence, a kernel is an image measure with an extra parameter ω, i. e. a parametrized family of image measures. Much like in functional analysis, a Markov kernel maps a measure μ on $\mathcal{A}_1$ to a new measure μ_2 on $\mathcal{A}_2$ by means of integration. For $A \in \mathcal{A}_2$,

$$\mu_2(A) := \int_{\Omega_1} K(\omega, A) d\mu(\omega)\ . \qquad \text{(B.0.14)}$$

When is a Conditional Probability Regular?

Regular conditional probabilities are obviously the kind of probability to work with preferrably. Since one is usually free to choose any version of a given conditional probability, the the crucial question is whether a given conditional probability has a regular version. This is not always true, but can be guaranteed on any space that is locally sufficiently similar to Euclidean space. This local similarity is formalized by the concept of a Borel space, and Th. 44 states that any conditional distribution on such a space has an equivalent regular version.

Appendix C

Dominated and Undominated Models

Probability models on infinite-dimensional spaces can raise a number of complications not usually familiar from the finite-dimensional, density-based case. One such complication are undominated families of models. In finite dimensions, all models in a given model family $\{\mu_\theta | \theta \in \Omega_\theta\}$ can usually be represented as densities with respect to one and the same reference measure (also called the *carrier measure*). The condition required for a measure to be representable as a density is simple and not very restrictive. Let μ_θ be any measure in the model family. Let $\text{Null}(\mu_\theta)$ be the system of its null sets, i. e. the set of all measurable sets A for which $\mu_\theta(A) = 0$. According to the Radon-Nikodym theorem, μ_θ has a density representation w. r. t. another, given measure ν if and only if $\text{Null}(\nu) \subset \text{Null}(\mu_\theta)$. Then there is some density function p_θ such that

$$d\mu_\theta = p_\theta d\nu \ . \tag{C.0.1}$$

That the null set condition is necessary for the existence of a density is perfectly intuitive: If $d\nu$ integrates to zero on a given set A, then it is not possible to reweight it by a density function p_θ such that $p_\theta d\nu$ integrates to a finite, non-zero value. That the condition is also sufficient may be a bit more surprising, because it implies that representability by a density function does not involve any kind of smoothness condition on the measures. All that is required is a condition on the null sets. This condition has a binary nature, because the magnitude of the non-zero values of ν is not relevant.

The null set condition is also referred to as absolute continuity of the measures: μ_θ is called *absolutely continuous* w. r. t. ν, in symbols $\mu_\theta \ll \nu$, if $\text{Null}(\nu) \subset \text{Null}(\mu_\theta)$. Representation of a family $\{\mu_\theta | \theta \in \Omega_\theta\}$ as a family of densities is convenient only if all measures μ_θ are absolutely continuous w. r. t. one and the same measure ν. Densities with respect to different carrier measures are not comparable. A suitable carrier measures is called a *dominating measure.*

Definition 57. Let $\mathcal{M}$ be a family of measures and ν any measure, not necessarily in $\mathcal{M}$. Then ν *dominates* $\mathcal{M}$ if $\mu \ll \nu$ for all $\mu \in \mathcal{M}$. If such any such measure ν exists for $\mathcal{M}$, the family is called *dominated.* If not, it is called *undominated.*

In Bayesian nonparametrics and other infinite-dimensional modeling problems, many interesting model families are not dominated.

Undominated Families: Intuition

A dominating measure ν must be finite, or at least σ-finite, to be meaningful for the definition of a density. Such a measure can only assign non-zero mass to a limited number of sets. Therefore, on large σ-algebras, measures must have a large number of null sets to be σ-finite. For example, a σ-finite measure on the real line cannot assign non-zero mass to all singletons. If it did, all non-empty open intervalls would have infinite measure. Undominated families exist because the absolute continuity condition for the existence of a density requires that only those sets may be null sets for the dominating measure which are also null sets for *all* measures in the family. If the family is too diverse, the number of null sets common to all its elements is not large enough to permit the definition of a dominating measure.

More precisely, a dominating measure for a family $\mathcal{M}$ must assign non-zero mass to all sets which are non-null under *any* measure in $\mathcal{M}$. Define the null system of $\mathcal{M}$ as

$$\text{Null}(\mathcal{M}) := \bigcap_{\mu \in \mathcal{M}} \text{Null}(\mu) \ . \tag{C.0.2}$$

In other words, ν dominates $\mathcal{M}$ if and only if $\text{Null}(\nu) \subset \text{Null}(\mathcal{M})$. If the family $\mathcal{M}$ is too large, and the null set patterns of the different measures in $\mathcal{M}$ are too diverse, then $\text{Null}(\mathcal{M})$ becomes too small for any σ-finite measure to satisfy $\text{Null}(\nu) \subset \text{Null}(\mathcal{M})$, and $\mathcal{M}$ is undominated.

Here is an example: For $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, denote by δ_x the Dirac measure centered at x, and by $G(x,1)$ the Gaussian measure of mean x and unit variance. First, consider the family of all Diracs $\mathcal{D} := \{\delta_x | x \in \mathbb{R}\}$. For every $x \in \mathbb{R}$, there is a Dirac which assigns non-zero mass to the singleton $\{x\}$, and hence $\text{Null}(\mathcal{D}) = \emptyset$. The family of Diracs is too large and diverse to be dominated. But the diversity is essential: The Gaussian family $\mathcal{N} := \{G(x,1)|x \in \mathbb{R}\}$ is just as large. But any two Gaussians $G(x_1,1)$ and $G(x_2,2)$ have exactly the same null sets, and so $\mathcal{N} := \{G(x,1)|x \in \mathbb{R}\}$ has null system $\text{Null}(\mathcal{N}) = \text{Null}(G(0,1))$. Since this is the null system of a σ-finite measure (the Gaussian), it is not too complicated to admit a dominating measure.

Criterion for Undominated Families

Proving that a family of measures is dominated may be an arbitrarily difficult endeavor, but the following lemma, due to Halmos and Savage (1949) provides a useful criterion for proving the converse. It uses the following terminology: Let $\mathcal{M}, \mathcal{N}$ be two sets of measures. The set $\mathcal{N}$ is said to be dominated by $\mathcal{M}$, denoted $\mathcal{N} \ll \mathcal{M}$, if every measure in $\mathcal{N}$ is dominated by *some* measure in $\mathcal{M}$. In particular, $\mathcal{N}$ being dominated by a set does not imply domination by a single measure. The sets are called equivalent, or $\mathcal{M} \sim \mathcal{N}$, if $\mathcal{M} \ll \mathcal{N}$ and $\mathcal{N} \ll \mathcal{M}$.

Lemma 58 (Halmos-Savage)**.** *If a set of measures is dominated (by a single measure), it has an equivalent countable subset.*

The result is non-trivial because, even if a set $\mathcal{M}$ is dominated by some measure ν, the dominating measure need not be contained in $\mathcal{M}$. If $\nu \in \mathcal{M}$, then trivially $\mathcal{M} \sim \{\nu\}$. An interesting aspect of the lemma is that it makes the inherent connection between domination and countability explicit: Equivalent measures have identical null sets, so intuitively, the lemma states that a dominated set can only have a countable number of different null set patterns. As an intermediate result in their proof, Halmos and Savage (1949) construct a dominating measure as a countable convex combination of measures in $\mathcal{M}$. The construction result is reproduced here as a corollary, and is used in the proof of La. 15.

Corollary 59. *If a set $\mathcal{M}$ of measures is dominated by a σ-finite measure, there exists a countable sequence $\{\mu_i\}_{i\in\mathbb{N}}$ of measures $\mu_i \in \mathcal{M}$, and a countable sequence of non-negative coefficients $\{c_i\}_{i\in\mathbb{N}}$ with $\sum_i c_i = 1$, such that the measure defined by*

$$\rho = \sum_{i\in\mathbb{N}} c_i \mu_i \tag{C.0.3}$$

dominates $\mathcal{M}$.

A consequence of relevance for Bayesian nonparametrics is the following:

Remark 60 (The DP posterior family on $\mathbb{R}$ is not dominated)**.** A Dirichlet process $\text{DP}(\alpha G_0)$ on the real line has a family of posteriors (under sample size one) defined by $\{\mathrm{DP}\,(\alpha G_0 + \delta_x)\,|x \in \mathbb{R}\}$. If G_0 is continuous, this family is undominated, which can be proven by means of La. 58. By the well-known posterior formula, the probability under the posterior $\mathrm{DP}\,(\alpha G_0 + \delta_{x_1})$ of a Dirac δ_{x_2} is non-zero, $\mathrm{DP}\,(\delta_{x_2}|\alpha G_0 + \delta_{x_1}) > 0$, if and only if $x_1 = x_2$. Therefore, $\mathrm{DP}\,(\alpha G_0 + \delta_{x_1}) \ll \mathrm{DP}\,(\alpha G_0 + \delta_{x_2})$ if and only if $x_1 = x_2$, so any dominating subset of the posterior family must contain $\mathrm{DP}\,(\alpha G_0 + \delta_x)$ for each $x \in \mathbb{R}$ and is therefore uncountable. Then by La. 58, $\{\mathrm{DP}\,(\alpha G_0 + \delta_x)\,|x \in \mathbb{R}\}$ cannot be dominated. The argument carries over the case of an arbitrary (but countable) number of observations, and to Dirichlet processes on any uncountable domain. In particular, the fact that the posterior family is not dominated implies that it is not dominated by the prior, such that there can be no closed-form substitute for the Bayesian formula.

Appendix D

Notation

Symbol	Description	Reference
Λ	Abstract probability space, i. e. the common domain of all random variables (the ranges of the random variables are the respective sample spaces)	Sec. 2.1
X	Random variable, in particular observation random variable	Sec. 2.1
Y	Random variable, in particular hyperparameter of a model	Sec. 2.1
Θ	Parameter random variable in a Bayesian model	Def. 1, p. 19
Ω_x	Sample space of random variable X	Sec. 2.1
$\mathcal{A}$	σ-algebra on abstract probability space Λ	Sec. 2.1
$\mathcal{C}$	Arbitrary σ-algebra	Sec. 2.1
$\mathcal{B}$	Borel σ-algebra	Sec. 2.1
$\mathbb{P}$	Abstract probability measure on $(\Lambda, \mathcal{A})$	Sec. 2.1
μ_X	Measure specifying distribution of random variable X, i. e. the image measure $\mu_X = X(\mathbb{P})$	Sec. 2.1
$p_{X\|\theta}$	Density of a conditional measure $\mu_{X\|\theta}$	Def. 2, p. 20
$\mu_{X\|\theta}$	Conditional measure of X given $\Theta = \theta$	App. B
$\mu(X\|\mathcal{C})$	Conditional distribution of X given a σ-algebra $\mathcal{C}$	App. B

Symbol	Description	Reference
$\sigma(X)$	σ-algebra generated by X ($\sigma(X) = X^{-1}(\mathcal{B}_x)$ if X takes values in $(\Omega_x, \mathcal{B}_x)$)	Def. 53, p. 192
$\frac{d\mu}{d\nu}$	Radon-Nikodym derivative (density) of μ w. r. t. ν	
$\ll$	absolute continuity relation on measures ($\mu \ll \nu$, μ is absolutely continuous w. r. t. ν)	App. C
ν-a.e.	Almost everywhere, with respect to measure ν	
X^{-1}	Inverse of a random variable (with the random variable regarded as a mapping)	
ω	"Atomic" or "elementary" event, an element of the abstract probability space Λ	
$\mathbb{E}[X]$	Expectation of random quantity X	
$\mathbb{E}_{\mu_{X\|y}}$	Expectation evaluated w. r. t. conditional measure $\mu_{X\|y}$	
$\mathbb{E}[X\|\mathcal{C}]$	Conditional expectation of X given σ-algebra $\mathcal{C}$	App. B
$\mathbb{E}[X\|Y]$	Conditional expectation $\mathbb{E}[X\|\mathcal{C}]$ with $\mathcal{C} = \sigma(Y)$	App. B
Z	(1) Partition function of a density	Sec. 2.2.3
	(2) Mixing variable (a random variable) in a mixture model	Sec. 2.5
$\mathcal{S}[p]$	Entropy (as a functional of the density p)	Sec. 2.2.3
$H(\theta)$	Potential function or energy function of probability distribution	Sec. 2.2.3
K	Number of clusters in a grouping problem (possibly a random quantity)	
N_{counts}	Number of counts in a histogram	
N_{bins}	Number of bins in a histogram (or Dirichlet random vector)	
n	Number of observations	
n_k	Number of observations assigned to a cluster with index k	
$\mathbf{h}$	Vector variable representing a histogram	
$\mathbb{S}(r)$	Symmetric group of order r	Sec. 3.1.2
π	(i) Permutation $\pi \in \mathbb{S}(r)$ (in context of rankings)	Sec. 3.1.2 Sec. 4.1.4
	(ii) Expectation parameter of a Dirichlet distribution	

Symbol	Description	Reference
$C(\pi)$	Consistent set of a partial ranking π (the set of all possible completions of π)	Sec. 3.1.2
$\mathrm{Sim}\left(\mathbb{R}, d\right)$	Standard simplex in $\mathbb{R}^d$	
$\mathrm{DP}\left(\alpha G_0\right)$	Dirichlet process with scatter parameter $\alpha \in \mathbb{R}_+$ and base measure G_0	Def. 25
δ_{ij}	Kronecker symbol	
$\delta_{x_0}(x)$	Dirac delta function or Dirac measure at x_0	
$\mathrm{Cov}\left[X, Y\right]$	Covariance of X and Y	
$\mathbb{I}_A$	Indicator function of set A	
E	Index set of a product space Ω^{E} (arbitrary set, usually infinite)	Sec. 2.4.2, 5.3
I, J	Index sets, usually subsets of E	Sec. 2.4.2, 5.3
E^*	Set of all finite subsets of E	Sec. 2.4.2, 5.3
Ω^{I}	Product space $\Omega^{\mathrm{I}} = \prod_{i \in I} \Omega$	Sec. 2.4.2, 5.3
$\mathcal{B}^{\mathrm{I}}$	Borel product algebra $\bigotimes_{i \in I} \mathcal{B}(\Omega)$ on Ω^{I}	Sec. 2.4.2, 5.3
X^{I}	Random variable with sample space $(\Omega^{\mathrm{I}}, \mathcal{B}^{\mathrm{I}})$	Sec. 2.4.2, 5.3
μ^{I}	Measure of X^{I} (*not* a product measure, in general)	Sec. 2.4.2, 5.3
$\mathrm{P}_{\mathrm{J,I}}$	Projection operator between product spaces $\Omega^{\mathrm{J}} \supset \Omega^{\mathrm{I}}$	Sec. 2.4.2, 5.3
$\mathrm{R}_{\mathrm{J,I}}$	Preimage under projection, $\mathrm{R}_{\mathrm{J,I}} = \mathrm{P}_{\mathrm{J,I}}^{-1}$	Sec. 2.4.2, 5.3
$\mathcal{H}(\mathcal{B})$	Set of $\mathcal{B}$-measurable partitions (partitions consisting of sets in $\mathcal{B}$)	Sec. 2.4.4
$\mathcal{H}(\mathcal{B})^*$	Set of $\mathcal{B}$-measurable partitions containing only a finite number of sets	Sec. 2.4.4
$\mathrm{P}_{\mathrm{J,I}}^*$	Projection operator on conditional probabilities	Def. 30
$\mathrm{proj\,lim}\,\mathcal{M}$	Projective limit (a measure) of a projective family $\mathcal{M}$ of measures	Sec. 2.4.2

Bibliography

Ailon, N., Charikar, M., and Newman, A. (2005). Aggregating inconsistent information: Ranking and clustering. In *ACM Symposium on the Theory of Computing*.

Aldous, D. J. (1985). Exchangeability and related topics. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII - 1983*, number 1117 in Lecture Notes in Mathematics, pages 1–198. Springer.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, **65**(331), 1248–1255.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric estimation. *Annals of Statistics*, **2**(6), 1152–1174.

Arslan, O., Constable, P. D. L., and Kent, J. T. (1993). Domains of convergence for the EM algorithm: a cautionary tale in a location estimation problem. *Statistics and Computing*, **3**, 103–108.

Bach, F. and Jordan, M. I. (2004). Learning spectral clustering. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing System 16*, pages 305–312. MIT Press.

Bahadur, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.*, **25**, 423–462.

Barankin, E. M. and Maitra, A. P. (1963). Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics. *Sankhya*, **25**, 217–244.

Barndorff-Nielsen, O. E. (1970). *Exponential families. Exact theory*. Number 19 in Various Publications Series. Matematisk Institut, Aarhus Universitet.

Barndorff-Nielsen, O. E. (1973). *Exponential families and conditioning*. University of Copenhagen.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. John Wiley & Sons.

Barndorff-Nielsen, O. E. and Pedersen, K. (1968). Sufficient data reduction and exponential families. *Mathematica Scandinavia*, **22**, 197–202.

Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Annals of Statistics*, **27**, 536–561.

Basu, D. and Ghosh, J. K. (1969). Sufficient statistics in sampling from a finite universe. In *Proc. 36th Session Internat. Statist. Inst.*, pages 850–859.

Bauer, H. (1996). *Probability Theory*. W. de Gruyter.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite Markov chains. *Annals of Mathematical Statistics*, **37**, 1554–1563.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

Beal, M. J., Ghrahmani, Z., and Rasmussen, C. (2002). The infinite hidden Markov model. In T. G. Dietterich, S. Becker, and Z. Ghrahmani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 577–584.

Beckett, L. A. (1993). Maximum likelihood estimation in Mallows' model using partially ranked data. In M. A. Fligner and J. S. Verducci, editors, *Probability Models and Statistical Analyses for Ranking Data*.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons.

Besag, J. (1974). Spatial interaction and the statistical anlysis of lattice systems. *Journal of the Royal Statistical Society*, **36**(2), 192–236.

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society*, **55**(1), 25–37.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, **10**(1), 3–66.

Bezdek, J. C., Hathaway, R. J., Howard, R. E., Wilson, C. A., and Windham, M. P. (1987). Local convergence analysis of a grouped variable version of coordinate descend. *Journal of Optimization Theory and Applications*, **54**, 471–477.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blackwell, D. (1973). Discreteness of Ferguson selections. *Annals of Statistics*, **1**(2), 356–358.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *Annals of Statistics*, **1**(2), 353–355.

Blackwell, D. and Ramamoorthi, R. V. (1982). A Bayes but not classically sufficient statistic. *Annals of Statistics*, **10**(3), 1025–1026.

Blei, D. M. (2004). *Probabilistic models for text and images*. Ph.D. thesis, U. C. Berkeley.

Blei, D. M. and Jordan, M. I. (2004). Variational methods for the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Bourbaki, N. (2003). *Topological Vector Spaces*. Springer.

Breckenridge, J. (1989). Replicating cluster analysis: Method, consistency and validity. *Multivariate Behavioral Research*, **24**, 147–161.

Bregler, C. (1997). Learning and recognizing human dynamics in video sequences. In *Proc. of IEEE CVPR 1997*.

Brown, L. (1964). Sufficient statistics in the case of independent random variables. *Annals of Mathematical Statistics*, **35**, 1456–1476.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families, with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics.

Bühlmann, H. (1960). *Austauschbare stochastische Variabeln und ihre Grenzwertsätze*. Ph.D. thesis. University of California Press, 1960.

Burkholder, D. L. (1961). Sufficiency in the undominated case. *Ann. Math. Statist.*, **32**, 1191–1200.

Busse, L. M., Orbanz, P., and Buhmann, J. M. (2007). Cluster analysis of heterogeneous rank data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 113–120.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, **98**, 1969–1987.

Cipra, B. A. (1987). An introduction to the Ising model. *American Mathematical Monthly*, **94**(10), 937–959.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, **91**(434), 883–904.

Crain, B. R. (1973). A note on density estimation using orthogonal expansions. *Journal of the American Statistical Association*, **68**, 964–965.

Crain, B. R. (1974). Estimation of distributions using orthogonal expansions. *Annals of Statistics*, **2**, 454–463.

Crain, B. R. (1976a). Exponential models, maximum likelihood estimation, and the haar condition. *Journal of the American Statistical Association*, **71**, 737–740.

Crain, B. R. (1976b). More on estimation of distributions using orthogonal expansions. *Journal of the American Statistical Association*, **71**, 741–745.

Critchlow, D. (1985). *Metric Methods for Analyzing Partially Ranked Data*. Springer.

Csiszár, I. and Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decision*, pages 205–237. Supplement Issue 1.

Dalal, S. R. and Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society B*, **45**(2), 278–286.

Darmois, G. (1935). Sur les lois de probabilite a estimation exhaustive. *C. R. Acad. Sci. Paris*, **260**, 1265–1266.

de Finetti, B. (1931). Fuzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei*, **4**, 251–299.

de Leeuw, J. (1994). Block relaxation algorithms in statistics. In H. H. Bock, W. Lenski, and M. M. Richter, editors, *Information Systems and Data Analysis*, pages 308–325. Springer.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analyis. *Journal of the American Society for Information Science*.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. John Wileay & Sons.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B*, **39**(1), 1–38.

Denny, J. L. (1964). On continuous sufficient statistics. *Annals of Mathematical Statistics*, **35**(3), 1229–1233.

Denny, J. L. (1967). Sufficient conditions for a family of probabilities to be exponential. *Proceedings of the National Academy of Sciences*, **57**, 1184–1188.

Denny, J. L. (1972). Sufficient statistics and discrete exponential families. *Annals of Mathematical Statistics*, **43**(4), 1320–1322.

Devroye, L. (1986). *Non-uniform random variate generation*. Springer.

Diaconis, P. (1988). *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics.

Diaconis, P. (1989). A generalization of spectral analysis with applications to ranked data. *Annals of Statistics*, **17**(3), 949–979.

Diaconis, P. and Freedman, D. (1980). DeFinetti's theorem for Markov Chains.

Diaconis, P. and Freedman, D. (1984). Partial exchangeability and sufficiency. In J. K. Ghosh and J. Roy, editors, *Statistics: Applications and New Directions*. Indian Statistical Institute.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *Annals of Statistics*, **14**(1), 1–67.

Diaconis, P. and Freedman, D. A. (1998). Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli*, **4**, 411–444.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, **7**(2), 269–281.

Doksum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability*, **2**, 183–201.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholdingdensity estimation by wavelet thresholding. *Annals of Statistics*, **24**, 508–539.

Doob, J. L. (1953). *Stochastic Processes*. J. Wiley & Sons.

Draper, D. (1999). Discussion of "bayesian nonparametric inference for random distributions and related functions", by s. g. walker, p. damien, p. w. laud, and a. f. m. smith. *Journal of the Royal Statistical Society*, **61**, 485–527.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, **3**(7).

Dynkin, E. B. (1961). Necessary and sufficient statistics for a family of probability distributions. *Selected Translations: Mathematical Statistics and Probability*, **1**, 23–41.

Efron, B. (1978). The geometry of exponential families. *Annals of Statistics*, **6**(2), 362–376.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **89**(425), 268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**(430), 577–588.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, volume II. Wiley.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**(2).

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**(4), 615–629.

Ferguson, T. S. and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Annals of Statistics*, **7**, 163–186.

Fiorot, J. C. and Huard, P. (1979). Composition and union of general algorithms of optimization. *Mathematical Programming Study*, **10**, 69–85.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A*, **222**, 309–368.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. London A*, **144**, 285–307.

Fligner, M. A. and Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society B*, **48**(3), 359–369.

Forsyth, D. A. and Ponce, J. (2003). *Computer Vision: a modern approach*. Prentice Hall.

Fremlin, D. H. (2003–2006). *Measure Theory*, volume I-IV. University of Essex.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman, D., Geman, S., Graffigne, C., and Dong, P. (1990). Boundary detection by constrained optimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **12**(7), 609–628.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **PAMI-6**(6), 721–741.

Gersho, A. and Gray, R. M. (1992). *Vector quantization and signal compression*. Kluwer Academic Publishing.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2002). Convergence rates of posterior distributions. *Annals of Statistics*, **28**(2), 500–531.

Ghosh, J. K. and Ramamoorthi, R. V. (2002). *Bayesian Nonparametrics*. Springer.

Ghosh, J. K., Morimoto, H., and Yamada, S. (1981). Neyman factorization and minimality of pairwise sufficient subfields. *Ann. Stat.*, **9**, 514–530.

Gikhman, I. I. and Skorohod, A. V. (1974). *The Theory of Stochastic Processes, Vols. I and II*. Springer.

Goldberger, J. and Greenspan, H. (2006). Context-based segmentation of image sequences. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **28**(3), 463–468.

Goodman, J. and Sokal, A. D. (1989). Multigrid Monte Carlo. Conceptual foundations. *Phys. Rev. D*, **40**(6), 2035–2071.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732.

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, **28**, 355–375.

Grenander, U. (1981). *Abstract inference*. Wiley.

Grenander, U. (1983). Tutorial in pattern recognition.

Griffin, J. E. (2007). The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference. Technical report, University of Kent.

Griffiths, T. L. and Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*.

Hall, P. (1986). On the rate of convergence of orthogonal series density estimation. *Journal of the Royal Statistical Society*, **48**, 115–122.

Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Stat.*, **20**, 225–241.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices.

Hastie, T. and Tibshirani, R. J. (1996). Disicriminant analysis of Gaussian mixtures. *Journal of the Royal Statistical Society*, **58**, 155–176.

Heiser, W. J. (1995). Convergent computing by iterative majorization: Theory and applications in multidimensional data analysis. In W. J. Krzanowski, editor, *Recent Advances in Descriptive Multivariate Analysis*, pages 157–189. Clarendon Press.

Hermes, L., Zöller, T., and Buhmann, J. M. (2002). Parametric distributional clustering for image segmentation. In *Computer Vision - ECCV '02*, volume 2352 of *LNCS*, pages 577–591. Springer.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, **80**(2), 470–501.

Higdon, D., Lee, H., and Holloman, C. (2003). Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems. In J. M. Bernardo et al., editor, *Bayesian Statistics 7*, pages 181–198.

Hipp, C. (1974). Sufficient statistics and exponential families. *Annals of Statistics*, **2**(6), 1283–1292.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, **18**(3), 1259–1294.

Hobart, G. L. and Jones, J. P. (2001). Honest exploration of intracatable probability distributions via Markov chain Monte Carlo. *Statistical Science*, **16**(4), 312–334.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 50–57.

Hofmann, T. and Buhmann, J. (1997). Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(1), 1–14.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *American Statistician*, **58**(1).

Huzurbazar, V. S. (1976). *Sufficient statistics*. Marcel Dekker.

Ionescu Tulcea, C. T. (1950). Mesures dans les espaces produits. *Atti Accad. Naz. Lincei Rend.*, **7**, 208–211.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, **31**(3), 264– 323.

Jain, S. and Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, **13**, 158–182.

Jain, S. and Neal, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, **2**, 445–472.

James, L. F., Lijoi, A., and Prünster, I. (2005). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. of Statistics*, **33**, 105–120.

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181–214.

Kallenberg, O. (1997). *Foundations of Modern Probability*. Springer.

Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81–93.

Kerov, S. V., Olshanski, G., and Vershik, A. M. (2004). Harmonic analysis on the infinite symmetric group. *Inventiones Mathematicae*, **158**, 551–642.

Khan, S. and Shah, M. (2001). Object based segmentation of video, using color, motion and spatial information. In *Proc. of IEEE CVPR 2001*.

Kingman, J. F. C. (1978). Uses of exchangeability. *Annals of Probability*, **6**, 183–197.

Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, **34**(2), 837–877.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer.

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability*. Chelsea Publ. Co.

Koopman, B. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, **39**, 399–409.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Mutltivariate Distributions, Vol. 1*. John Wiley & Sons, second edition.

Küchler, U. (1982a). Exponential families of Markov processes, Part I: General results. *Math. Oper. Statist., Ser. Statist.*, **13**, 57–69.

Küchler, U. (1982b). Exponential families of Markov processes, Part II: Birth-and-death processes. *Math. Oper. Statist., Ser. Statist.*, **13**, 219–230.

Küchler, U. and Sørensen, M. (1997). *Exponential Families of Stochastic Processes*. Springer.

Lange, T., Roth, V., Braun, M., and Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Computation*, **16**(6), 1299–1323.

Lauritzen, S. L. (1988). *Extremal families and systems of sufficient statistics*, volume 49 of *LNS*. Springer.

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *Canadian Journal of Statistics*, **20**, 451–461.

Lebanon, G. and Lafferty, J. (2002). Cranking: Combining rankings using conditional probability models on permutations. In *International Conference on Machine Learning*.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, 3rd edition.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, **37**, 145–151.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computation with applications to a gene-regulation problem. *Journal of the American Statistical Association*, **89**(427), 958–966.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer.

Liu, J. S. and Sabatti, C. (1998). Simulated sintering: Markov chain Monte Carlo with spaces of varying dimensions. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 386–413.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes. *Biometrika*, **81**(1), 27–40.

Loève, M. (1977a). *Probability Theory*, volume 1. Springer.

Loève, M. (1977b). *Probability Theory*, volume 2. Springer.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, **23**, 727–741.

Mallows, C. L. (1957). Non-null ranking models I. *Biometrika*, **44**, 114–130.

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. Chapman & Hall.

Martin, D., Fowlkes, C., and Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**(5), 530–549.

Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992). Polya trees and random distributions. *Annals of Statistics*, **20**, 1203–1221.

McAuliffe, J. D., Blei, D. M., and Jordan, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*, **16**, 5–14.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

McLachlan, G. J. and Gordon, R. D. (1989). Mixture models for partially unclassified data: a case study of renal venous renin levels in essential hypertension. *Statistics in Medicine*, **8**, 1291–1300.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons.

Meilă, M. and Bao, L. (2008). Estimation and clustering with infinite rankings. In D. McAllester and P. Millimäki, editors, *Proceedings of the 24-th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*.

Mengersen, K. L. and Robert, C. P. (1995). Testing for mixture via entropy distance and Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 5*.

Moh, Y., Orbanz, P., and Buhmann, J. M. (2008). Music preference learning with partial information. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. In press.

Morel, J.-M. and Solimini, S. (1995). *Variational Methods for Image Segmentation*. Birkhäuser.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, **19**(1), 95–111.

Murphy, T. B. and Martin, D. (2003). Mixtures of distance-based models for ranking data. *Computational Statistics and Data Analysis*, **41**, 645–655.

Neal, R. M. (1991). Bayesian mixture modeling by Monte Carlo simulation. Technical report, Department of Computer Science, University of Toronto.

Neal, R. M. (1994). *Bayesian Learning for Neural Networks*. Ph.D. thesis.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Neal, R. M. (2003). Density modeling and clustering using dirichlet diffusion trees. In J. M. Bernardo et al., editor, *Bayesian Statistics 7*, pages 619–629.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, **8**(4), 343–366.

Neyman, J. (1935). Su un teorema concernente le cosiddette statistiche sufficienti. *Inst. Ital. Atti. Giorn.*, **6**, 320–334.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge Univserity Press.

O'Hagan, A. (1991). Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, **29**, 245–260.

Øksendal, B. (1992). *Stochastic Differential Equations*. Springer.

Oliver, C. and Quegan, S. (1998). *Understanding Synthetic Aperture Radar Images*. Artech House.

Olshanski, G. (2003). Point processes related to the infinite symmetric group. In V. O. Ch. Duval, L. Guieu, editor, *The orbit method in geometry and physics: in honor of A.A. Kirillov*, pages 349–393.

Orbanz, P. and Buhmann, J. M. (2005). SAR images as mixtures of Gaussian mixtures. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 2, pages 209–212.

Orbanz, P. and Buhmann, J. M. (2006). Smooth image segmentation by nonparametric Bayesian inference. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 444–457. Springer.

Orbanz, P. and Buhmann, J. M. (2007). Bayesian order-adaptive clustering for video segmentation. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, pages 334–349. Springer.

Orbanz, P. and Buhmann, J. M. (2008). Nonparametric Bayesian image segmentation. *Int. J. of Computer Vision.*, **77**(1–3), 25–45.

Parthasarathy, K. R. (1967). *Probability measures on metric spaces*. Academic Press.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc.*, **185**, 71–110.

Pillai, N. S., Wu, Q., Liang, F., Mukherjee, S., and Wolpert, R. L. (2007). Characterizing the function space for Bayesian kernel models. *Journal of Machine Learning Research*, **8**, 1769–1797.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophical Society*, **32**, 567–579.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, **102**, 145–158.

Puzicha, J., Hofmann, T., and Buhmann, J. M. (1999). Histogram clustering for unsupervised segmentation and image retrieval. *Pattern Recognition Letters*, **20**, 899–909.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard University Press.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In T. K. L. S. A. Solla and K. R. Mller, editors, *Advances in Neural Information Processing Systems 12*, pages 554–560.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society*, **59**(4), 731–792.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, **11**, 416–431.

Robert, C. P. (1995). Mixtures of distributions: inference and estimation. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parametrization for the Gibbs sampler. *Journal of the Royal Statistical Society*, **59**, 291–317.

Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression and related optimization problems. *Proc. of the IEEE*, **86**(11), 2210–2239.

Roth, V., Laub, J., Motoaki, K., and Buhmann, J. M. (2003). Optimal cluster preserving embedding of non-metric proximity data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(12), 1540–1551.

Ryll-Nardzewski, C. (1957). On stationary sequences of random variables and the de Finetti's [sic] theorem. *Colloquium Mathematicum*, **4**, 149–156.

Samson, C., Blanc-Féraud, L., Aubert, G., and Zerubia, J. (2000). A variational model for image classification and restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(5), 460–472.

Schervish, M. and Carlin, B. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics*, **1**, 111–127.

Schervish, M. J. (1995). *Theory of Statistics*. Springer.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics*, **29**, 687–714.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 888–905.

Skorohod, A. V. (1974). *Integration in Hilbert space*. Springer.

Socci, N. D., Lee, D. D., and Seung, H. S. (1998). The rectified gaussian distribution. In M. I. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Proscessing Systems*, pages 350–356. MIT Press.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Annals of Statistics*, **14**(1), 138–150.

Stoica, P. and Selen, Y. (2004). Model order selection: A review of information criterion rules. *IEEE Signal Processing*, pages 36–47.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2006). Describing visual scenes using transformed Dirichlet processes. In *Advances in Neural Information Processing Systems*.

Sudderth, E. B. (2006). *Graphical Models for Visual Object Recognition and Tracking*. Ph.D. thesis.

Susarla, V. and Ryzin, J. V. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71**(356), 897–902.

Swendsen, R. H. and Wang, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulation. *Physical Review Letters*, **58**, 86–88.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions. *Journal of the American Statistical Association*, **82**, 528–550.

Teh, Y. W. (2006). A Bayesian interpretation of interpolated Kneser-Ney. Technical report, School of Computing, National University of Singapore.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, volume 17.

Teh, Y. W., Daumé III, H., and Roy, D. M. (2008a). Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems*.

Teh, Y. W., Kurihara, K., and Welling, M. (2008b). Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems*. To appear.

Tekalp, A. M. (2000). Video segmentation. In A. C. Bovik, editor, *Handbook of Image and Video Processing*, pages 383–399.

Thibaux, R. and Jordan, M. I. (2006). Hierarchical beta processes and the Indian buffet process. In *Uncertainty in Artificial Intelligence*.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**(4), 1701–1762.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1995). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons.

Tu, Z. and Zhu, S.-C. (2002). Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(5), 657–673.

van de Geer, S. A. (1993). Hellinger consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics*, **21**, 14–44.

Wainwright, M. J. and Jordan, M. I. (2003). Graphical models, exponential families, and variational inference. Technical Report 649, University of California, Berkeley.

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society B*, **61**(3), 485–527.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.

Weiss, Y. and Adelson, E. H. (1996). A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *Proc. of IEEE CVPR 1996*.

West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*. Springer.

Willsky, A. S. (2002). Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, **90**(8), 1396–1458.

Winkler, G. (2003). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer.

Wolpert, R. L. and Ickstadt, K. (2004). Reflecting uncertainty in inverse problems: A Bayesian solution using Lévy processes. *Inverse Problems*, **20**(6), 1759–1771.

Wolpert, R. L., Ickstadt, K., and Hansen, M. B. (2003). *A nonparametric Bayesian appraoch to inverse problems (with discussion)*, pages 403–418.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, **11**, 95–103.

Xie, H., Pierce, L. E., and Ulaby, F. T. (2002). Statistical properties of logarithmically transformed speckle. *IEEE Trans. on Geoscience and Remote Sensing*, **40**(3).

Yamasaki, Y. (1985). *Measures on infinite dimensional spaces*. World Scientific Publishing.

Zaragoza, H., Hiemstra, D., Tipping, D., and Robertson, S. (2003). Bayesian extension to the language model for ad hoc information retrieval. In *Proc. SIGIR 2003*.

Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *Annals of Statistics*, **28**, 532–552.