


Kernel-based learning of orthogonal functions

Journal Article**Author(s):**

Scampicchio, Anna ; Bisiacco, Mauro; Pillonetto, Gianluigi

Publication date:

2023-08-07

Permanent link:

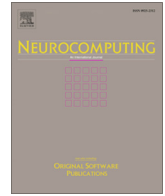
<https://doi.org/10.3929/ethz-b-000614385>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

Neurocomputing 545, <https://doi.org/10.1016/j.neucom.2023.126237>



Kernel-based learning of orthogonal functions

Anna Scampicchio^{a,*}, Mauro Bisiacco^b, Gianluigi Pillonetto^b

^a Institute for Dynamic Systems and Control, ETH, Sonneggstrasse 3, 8092 Zürich, Switzerland

^b Department of Information Engineering, University of Padova, via G. Gradenigo 6, Padova 35131, Italy

ARTICLE INFO

Article history:

Received 25 November 2021

Revised 5 November 2022

Accepted 22 April 2023

Available online 6 May 2023

Communicated by Zidong Wang

Keywords:

Function estimation

Orthogonality constraints

Kernel-based methods

Markov chain Monte Carlo

ABSTRACT

Estimating a set of orthogonal functions from a finite set of noisy data plays a crucial role in several areas such as imaging, dictionary learning and compressed sensing. The problem turns out especially hard due to its intrinsic non-convexity. In this paper, we solve it by recasting it in the framework of multi-task learning in Hilbert spaces, where orthogonality plays a role as inductive bias. Two perspectives are analyzed. The first one is mainly theoretic. It considers a formulation of the problem where non-orthogonal function estimates are seen as noisy data belonging to an infinite-dimensional space from which orthogonal functions have to be reconstructed. We then provide results concerning the existence and the convergence of the optimizers. The second one is more oriented towards applications. It consists in a learning scheme where orthogonal functions are directly inferred from a finite amount of noisy data. It relies on regularization in reproducing kernel Hilbert spaces and on the introduction of special penalty terms promoting orthogonality among tasks. The problem is then cast in a Bayesian framework, overcoming non-convexity through an efficient Markov chain Monte Carlo scheme. If orthogonality is not certain, our scheme can also understand from data if such form of task interaction really holds.

© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Machine learning applications typically require estimation of an unknown function from sparse and noisy data [1]. Regularization theory is often exploited to face this problem, and kernel-based methods play an important role in this context [2]. They connect Tikhonov regularization with reproducing kernel Hilbert spaces (RKHSs) [3,4], leading to important estimators like regularization networks and support vector machines [5].

A more complex scenario emerges when one has to reconstruct multiple functions known to share some common features, a problem often referred to as multi-task learning in the literature [6–9]. In this set-up, data related to a function can provide information also on the other ones, and this can be leveraged to improve the estimation performance. However, such dependencies among tasks have to be properly modelled. To this aim, RKHSs can be still adopted, in particular exploiting versions containing vector-valued functions [10] that define multi-task regularized kernel methods [11,12]. The unknown functions are then obtained by optimizing convex objectives that trade-off data fit and the information provided by the kernel, including tasks interactions.

The focus of this paper is a particular joint estimation problem that involves functions known to be mutually orthogonal. Notably, the problem is much more difficult than classic multi-task learning, since the presence of orthogonality constraints require estimators based on non-convex objectives.

1.1. Related work

Multi-task learning The motivation for multi-task learning stems from situations in which data are collected from different, yet related, experimental set-ups. Examples can be traced, e.g., in biomedical imaging [13], hospitalization management [14], inverse dynamics learning in robotics [15], and panel data analysis in econometrics [16, Part IV]. The core idea is to leverage task relatedness to enhance the estimation performance, with particular impact in situations when single-task data-sets are small: see, e.g., [17–19].

Many ways have been studied in the literature to encode task relatedness. A popular choice consists in (linear) mixed-effects models [20,21], where each function consists of the sum of a common term and a task-specific one, a model especially useful in population studies and pharmacokinetics [22–24]; see also [25] for a solution leveraging Markov Chain Monte Carlo (MCMC) within the Bayesian framework, and [26,27] for a non-parametric version using Gaussian process regression. Advantages of these

* Corresponding author.

E-mail addresses: ascampicc@ethz.ch (A. Scampicchio), bisiacco@dei.unipd.it (M. Bisiacco), giapi@dei.unipd.it (G. Pillonetto).

approaches, in different fields like biomedicine and imaging, are also documented, e.g., in [28–30].

Orthogonality constraints Joint estimation problems involving such a bias have been studied only for finite-dimensional spaces, and can be recast in the framework of Procrustes problems [31]. They find many important applications, e.g., in imaging [32], compressed sensing [33], factor analysis in psychometry [34] and object detection [35,36], dictionary learning [37], and conformal mapping [38]. Even if many sophisticated optimization algorithms have been proposed in the literature to address the intrinsic non-convexity of the orthogonality constraint [39–42], all of them can only guarantee convergence to local minima. An exception can be found in [43], where interesting advances are reported but restricted to the case of one or two active constraints.

1.2. Contribution

We recast function estimation subject to orthogonality constraints as a particular instance of kernel-based multi-task learning. In particular, the problem is first formalized by introducing RKHS norms and other special regularizers that promote orthogonality among the tasks. Such a formulation finds a connection with unbalanced orthogonal Procrustes problems [44], which in the particular case of interest do not admit an analytical solution (see, e.g., [40] Section 3.5.2). Then, we prove that our proposed formulation admits a stochastic interpretation, and describe it using Bayesian networks where the constraints derive from particular a priori probability density functions. We will see that any distribution of the single task, conditional on the data and all the remaining functions, preserves Gaussianity. Building upon this fact, non-convexity can be overcome by adopting MCMC in place of deterministic optimization [45]. In particular, we design a Gibbs sampling scheme able to efficiently reconstruct in sampled form the joint posterior of all the unknown tasks. Such a set-up allows also to learn from data the values of the hyper-parameters that regulate both function smoothness and the interaction among the tasks. This is useful in those circumstances where orthogonality is not certain. In fact, the proposed algorithm can also detect from data if some of the orthogonality constraints are not active and remove them from the estimation process.

To further investigate the problem of orthogonal functions estimation within RKHSs, we consider the following viewpoint. One can start considering non-orthogonal function estimates obtained by first exploiting decoupled estimators like regularization networks or support vector machines. Such estimates can be seen as infinite-dimensional noisy data and orthogonal estimates have then to be obtained by solving a non-convex, infinite-dimensional problem that incorporates the orthogonality constraints. Within this set-up, our contribution consists in proving the existence of the optimizers and their convergence when the data-set size (used to achieve the non-orthogonal estimates) grows to infinity.

1.3. Outline

The paper is organized as follows. In Section 2 the problem of orthogonal multi-task learning is formally introduced. In Section 3 a brief review of function estimation in the deterministic RKHS framework is first presented; then, we extend the approach to the orthogonal multi-task scenario, also providing its Bayesian interpretation. Such a probabilistic set-up is then exploited to design a Markov Chain Monte Carlo scheme that overcomes the non-convexity of the deterministic problem formulation. Section 4 is devoted to the theoretical analysis of orthogonal multi-task learning; the problem is faced assuming that data live in an infinite-dimensional space and proving existence and convergence

of the optimizers. Section 5 collects some numerical experiments that involve a large set of orthogonal and non-orthogonal tasks, and show the effectiveness of the computational scheme proposed in Section 3. Conclusions then end the paper.

2. Problem statement

The unknown r tasks (functions) are denoted by $f_i : \mathcal{X} \rightarrow \mathbb{R}$ for $i = 1, \dots, r$. The f_i belong to a Hilbert space \mathcal{H} with inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and are assumed mutually orthogonal, i.e.

$$\langle f_i, f_j \rangle_{\mathcal{H}} = 0, \quad i \neq j.$$

We can also think of a single unknown multi-task function that embeds all the f_i . It is denoted by $f : \mathcal{X} \times \{1, 2, \dots, r\}$ and belongs to the Hilbert space \mathcal{H}^r . Picking two arbitrary vector-valued functions g, h belonging to \mathcal{H}^r , then the inner product is given by

$$\langle g, h \rangle_{\mathcal{H}^r} = \sum_{i=1}^r \langle g_i, h_i \rangle_{\mathcal{H}}.$$

We assume that only a finite number of direct and noisy samples of each component of f is available. In particular, n_i input/output pairs $\{(\mathcal{X}_{ki}, \mathcal{Y}_{ki})\}_{k=1}^{n_i}$ are collected for each $i = 1, \dots, r$ and we use \mathcal{X}_i and \mathcal{Y}_i to indicate the sets of inputs and outputs for the i -th task. Then, for each f_i , the measurements model is

$$\mathcal{Y}_{ki} = f_i(\mathcal{X}_{ki}) + e_{ki}, \quad k = 1, \dots, n_i, \quad (1)$$

where all the noises e_{ki} are zero-mean independent Gaussians of variance σ_{ki}^2 . The problem is to estimate f from the data (1) exploiting also the inter-task information encoded in the orthogonal constraints.

3. Methodology

3.1. Single-task case: review of nonparametric estimation

Before delving into multi-task learning, we first recall the baseline nonparametric strategy for estimating a single f_i from $(\mathcal{X}_i, \mathcal{Y}_i)$ using Tikhonov regularization in a Reproducing Kernel Hilbert Space (RKHS) [3,46]. It consists in solving the program

$$\hat{f}_i = \arg \min_{f_i \in \mathcal{H}} \sum_{k=1}^{n_i} \frac{(\mathcal{Y}_{ki} - f_i(\mathcal{X}_{ki}))^2}{\sigma_{ki}^2} + \gamma_i \|f_i\|_{\mathcal{H}}^2, \quad (2)$$

where \mathcal{H} is a RKHS. This is a Hilbert space in which the evaluation functionals returning the value of f_i at a point $x \in \mathcal{X}$, defined as $\mathcal{E}_x f_i = f_i(x)$, are linear and continuous for any $x \in \mathcal{X}_i$. From Moore-Aronszajn theorem [3] it follows that each RKHS is in one-to-one correspondence with a positive semi-definite kernel operator

$$\mathcal{K} : \mathcal{X}_i \times \mathcal{X}_i \rightarrow \mathbb{R} \quad (3)$$

such that the so-called reproducing property for function evaluation reads as

$$\mathcal{E}_x f_i = f_i(x) = \langle \mathcal{K}(x, \cdot), f_i \rangle_{\mathcal{H}}.$$

These facts, together with Riesz-Frechet theorem (see, e.g., [47] [Chapter V, Theorem 1] or [48] [Theorem 6.19]) lead to the so called Representer Theorem [49]. It states that the solution of (2) has the structure of a regularization network, being a linear combination of kernel sections $\{\mathcal{K}(\mathcal{X}_{ki}, \cdot)\}_{k=1}^{n_i}$ centered at input locations in \mathcal{X}_i . The use of the quadratic loss in (2) to measure the adherence to data also implies that the expansion coefficients \hat{c}_i solve a linear system of equations. In particular, introducing

$$\Sigma_i = \text{diag}(\sigma_{1i}^2 \cdots \sigma_{n_i i}^2),$$

computing from (3) the kernel matrix $K_{\mathcal{H}} \in \mathbb{R}^{n_i \times n_i}$ as

$$[K_{\mathcal{H}}]_{a,b} = \mathcal{K}(x_a, x_b), \quad a, b = 1, \dots, n_i, \quad (4)$$

and collecting all outputs in the column vector Y_i , we get

$$\begin{aligned} \hat{c}_i &= \arg \min_{c_i} (Y_i - K_{\mathcal{H}} c_i)^\top \Sigma_i^{-1} (Y_i - K_{\mathcal{H}} c_i) + \gamma_i c_i^\top K_{\mathcal{H}} c_i \\ &= (K_{\mathcal{H}} + \gamma_i \Sigma_i)^{-1} Y_i. \end{aligned} \quad (5)$$

This result can be also given a Bayesian interpretation [50]. Indeed, let c_i be a Gaussian vector with prior distribution

$$c_i \sim \mathcal{N}(0, (\gamma_i K_{\mathcal{H}})^{-1}).$$

Next, consider the measurements model

$$Y_i = K_{\mathcal{H}} c_i + e_i$$

with the Gaussian noises

$$e_i \sim \mathcal{N}(0, \Sigma_i)$$

independent of c_i . Then, using standard results on estimation of Gaussian processes [51], one obtains that the minimum variance estimate \hat{c}_i of c_i given the data Y_i is

$$\hat{c}_i = (K_{\mathcal{H}} \Sigma_i^{-1} K_{\mathcal{H}} + \gamma_i K_{\mathcal{H}})^{-1} K_{\mathcal{H}} \Sigma_i^{-1} Y_i$$

and coincides with the estimator (5) obtained in the deterministic setting. More in general, the Bayesian interpretation of any optimization program is obtained if its objective can be interpreted as the negative logarithm of the likelihood-prior product of a suitable probabilistic model. Indeed, taking the exponential of the minus objective reported in (5), the Gaussian priors for c_i and e_i become immediately evident. This fact will be also exploited later on to deal with more complex regularized programs.

The Bayesian viewpoint is useful also when Σ_i and γ_i are unknown and have to be estimated from data as well. As described, e.g., in [52], one can interpret the prior variances of the functions and of the noises as random variables and assign them, e.g., non-informative priors that include (in practice) only non-negativity information. Then, stochastic simulation schemes relying on Markov Chain Monte Carlo [53] can reconstruct in sampled form the joint posterior of the function and the unknown variances. Minimum variance estimates of all the quantities of interest can then be computed by Monte Carlo integration.

3.2. Multi-task case: deterministic viewpoint

So far we have introduced two different Hilbert spaces. According to the problem statement reported in Section 2, our unknown tasks belong to the Hilbert space \mathcal{H} and they are orthogonal according to the metric induced by its inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A key example is given by the classical Lebesgue space of squared summable functions, i.e. $\mathcal{H} = \mathcal{L}^2$ with inner product $\langle f_i, f_j \rangle_{\mathcal{H}} = \int_0^1 f_i(x) f_j(x) dx$. In the previous subsection, we have then reviewed regularization networks where the tasks lie in a RKHS \mathcal{H} . This assumption is important since the estimator can embed information on function smoothness, e.g., continuous kernels like the popular spline or Gaussian induce spaces of continuous and differentiable functions [2].

Now, we face the orthogonal multi-task learning problem assuming that the functions belong to the intersection between the two Hilbert spaces \mathcal{H} and \mathcal{H} , using the two related metrics to regularize the problem. In particular, along the lines of (2), the general problem of multi-task orthogonal function estimation is stated as follows.

$$\begin{aligned} \hat{f} &= \arg \min_f \sum_{i=1}^r \left[\sum_{k=1}^{n_i} \frac{(y_{ki} - f_i(x_{ki}))^2}{\sigma_{ki}^2} + \gamma_i \|f_i\|_{\mathcal{H}}^2 \right] \\ &+ \sum_{i=1}^r \sum_{j>i} \gamma_{ij} |\langle f_i, f_j \rangle_{\mathcal{H}}|^2. \end{aligned} \quad (6)$$

Hence, the solution is the trade-off between data fit and two regularization terms. The first penalty is defined by the norm in \mathcal{H} and embeds smoothness. It acts on each task independently and is weighted by γ_i for $i = 1, \dots, r$ as already discussed in Section 3.1. The other term is defined by the inner-products in \mathcal{H} and is connected with the orthogonality constraints. The inter-task relations are tuned by γ_{ij} for $i = 1, \dots, r$ and $j > i$ (that avoids repetitions since the case $j < i$ is already taken into account by symmetry). Large values for γ_{ij} aim at setting $|\langle f_i, f_j \rangle_{\mathcal{H}}|$ as close to zero as possible. Clearly, for problem (6) to be well-defined, we assume that \mathcal{H} and \mathcal{H} have non-empty intersection. This in practice holds for all the situations of interest if the function domain \mathcal{X} is compact: the RKHSs used in practice contain only continuous functions that in turn are contained, e.g., in $\mathcal{H} = \mathcal{L}^2$.

Now, for computational reasons, we formulate a finite-dimensional approximation of (6). This is obtained by drawing inspiration from the representer theorem mentioned in Section 3.1, and also by the fact that any function in \mathcal{H} is given by a (possibly infinite) sum of kernel sections [4]. Specifically, we formulate each task as a linear finite combination of kernel functions over the input locations contained in the union of the \mathcal{X}_i whose cardinality is denoted by n . To simplify notation, we assume that $\mathcal{X}_1 = \dots = \mathcal{X}_r = \mathcal{X}$ and that such a set covers sufficiently well the domain \mathcal{X} . This assumption is stated without loss of generality: if for some i and input location x_k the output measurement y_{ki} is not available, we set the corresponding noise variance σ_{ki}^2 to infinity. Also, if there are regions of \mathcal{X} not well covered, we can add virtual input locations associated to measurements of infinite variance.

Assuming that the input location set is unique for all tasks implies that the $n \times n$ kernel matrices $K_{\mathcal{H}}$ defined by (4) are the same for each f_i . We also introduce the $n \times n$ Gram matrices $K_{\mathcal{H}}$ defined by the inner-product in \mathcal{H} as follows

$$[K_{\mathcal{H}}]_{a,b} = \langle \mathcal{K}(x_a, \cdot), \mathcal{K}(x_b, \cdot) \rangle_{\mathcal{H}}, \quad a, b = 1, \dots, n. \quad (7)$$

The two matrices $K_{\mathcal{H}}$ and $K_{\mathcal{H}}$ then permit to formulate the finite-dimensional approximation of (6), which is

$$\begin{aligned} \arg \min_{c_i} \sum_{i=1}^r & \left[(Y_i - K_{\mathcal{H}} c_i)^\top \Sigma_i^{-1} (Y_i - K_{\mathcal{H}} c_i) + \gamma_i c_i^\top K_{\mathcal{H}} c_i \right] \\ & i = 1 \dots r \\ & + \sum_{i,j>i} \gamma_{ij} c_i^\top K_{\mathcal{H}} c_j c_j^\top K_{\mathcal{H}} c_i. \end{aligned} \quad (8)$$

3.3. Multi-task case: Bayesian interpretation

Even if problem (8) is finite-dimensional, using this estimator remains difficult since the objective is non-convex. In addition, it contains unknown hyper-parameters: Σ_i, γ_i , and possibly also the interactions parameters γ_{ij} , have to be estimated from data as well. To solve this problem, below we introduce the Bayesian interpretation of (8). In this way, non convex optimization will be replaced by the problem of reconstructing in sampled form the tasks posterior by means of MCMC.

The first step is to find a suitable prior distribution for each c_i so that the objective in (8) can be interpreted as the negative logarithm of

$$p(Y_1, \dots, Y_r | c_1, \dots, c_r) p(c_1, \dots, c_r)$$

given all the hyper-parameters. The crux underlying the stochastic interpretation is related to the coupling of c_i and c_j . In this regard, we introduce the following fictitious model:

$$z_{ij} = \sqrt{c_i^\top K_{\mathcal{H}} c_j c_j^\top K_{\mathcal{H}} c_i} + \epsilon_{ij}, \quad i = 1, \dots, r, \quad j > i. \quad (9)$$

Then, we model ϵ_{ij} as zero-mean independent Gaussian noises of variance γ_{ij}^{-1} . Moreover, we assume that all the hyper-parameters γ_i, γ_{ij} and $\sigma_{k_i}^2$ for $i = 1, \dots, r, j > i, k = 1, \dots, n$ are mutually independent random variables (independent also of the noises). The resulting Bayesian network is presented in Fig. 1.

The joint density associated with such a model is the following. For ease of notation, let $Y = \{Y_i\}_{i=1}^r, Z = \{z_{ij}\}_{i=1, j>i}^r, c = \{c_i\}_{i=1}^r, \Sigma = \{\Sigma_i\}_{i=1}^r, \gamma = \{\gamma_i\}_{i=1}^r$ and $\Gamma = \{\gamma_{ij}\}_{i=1, j>i}^r$: then, the network architecture yields

$$\begin{aligned} p(Y, Z, c, \Sigma, \gamma, \Gamma) &= p(Y, Z|c, \Sigma, \gamma, \Gamma) p(c|\Sigma, \gamma, \Gamma) p(\Sigma) p(\gamma) p(\Gamma) \\ &= p(Y|c, \Sigma) p(Z|c, \Gamma) p(c|\gamma) p(\Sigma) p(\gamma) p(\Gamma) \\ &= \left(\prod_{i=1}^r p(Y_i|c_i, \Sigma_i) p(c_i|\gamma_i) p(\Sigma_i) p(\gamma_i) \right) \\ &\quad \times \left(\prod_{\substack{i=1 \\ j>i}}^r p(z_{ij}|c_i, c_j, \gamma_{ij}) p(\gamma_{ij}) \right). \end{aligned} \quad (10)$$

Now assume that hyper-parameters contained in Σ, γ and Γ are known and focus on the joint distribution $p(Y, Z, c|\Sigma, \gamma, \Gamma) = p(Y|c, \Sigma) p(Z|c, \Gamma) p(c|\gamma)$. Considering its factors, by (1) and (9) we have that

$$p(Y_i|c_i, \Sigma_i) \propto e^{-(Y_i - K_{\mathcal{H}} c_i)^\top \Sigma_i^{-1} (Y_i - K_{\mathcal{H}} c_i)} \quad (11a)$$

$$p(c_i|\gamma_i) \propto e^{-\gamma_i c_i^\top K_{\mathcal{H}} c_i} \quad (11b)$$

$$p(z_{ij}|c_i, c_j, \gamma_{ij}) \propto e^{-\gamma_{ij} \left(z_{ij} - \sqrt{c_i^\top K_{\mathcal{H}} c_j c_j^\top K_{\mathcal{H}} c_i} \right)^2}. \quad (11c)$$

The idea is to exploit the fictitious model (9) so that the negative logarithm of $p(Y, Z, c|\Sigma, \gamma, \Gamma)$ becomes proportional to the objective contained in (8). It is easy to see by inspection that this holds when all the virtual observations z_{ij} are set to zero.

Now, we investigate how the normal prior on the expansion coefficients c_i is influenced by the event $Z = 0$ and by the knowledge of all the other tasks and hyper-parameters. From (11b) and (11c), it is not difficult to see that such a conditional posterior remains Gaussian: indeed, it holds that

$$c_i | \gamma_i, c_{j \neq i}, \Gamma, Z = 0 \sim \mathcal{N}(0, P_i^{-1}), \quad \text{with}$$

$$P_i = \left(\gamma_i K_{\mathcal{H}} + \sum_{j>i} \gamma_{ij} K_{\mathcal{H}} c_j c_j^\top K_{\mathcal{H}} + \sum_{j<i} \gamma_{ji} K_{\mathcal{H}} c_j c_j^\top K_{\mathcal{H}} \right).$$

One can thus see that the inverse covariance of c_i , denoted by P_i , is given by the inverse of the unconditional covariance, i.e. $\gamma_i K_{\mathcal{H}}$, plus other terms that derive from the tasks interactions due to the possible orthogonality. Starting from this updated prior for c_i , we can now condition also on the measurements Y_i . Due to the Gaussianity, we easily obtain the following result

$$c_i | Y_i, z_{ij} = 0, c_{j \neq i}, \gamma_i, \Gamma, \Sigma_i \sim \mathcal{N}(\hat{c}_i, \hat{P}_i) \quad (12)$$

$$\begin{cases} \hat{c}_i = P_i^{-1} K_{\mathcal{H}} (K_{\mathcal{H}} P_i^{-1} K_{\mathcal{H}} + \Sigma_i)^{-1} Y_i \\ \hat{P}_i = (K_{\mathcal{H}} \Sigma_i^{-1} K_{\mathcal{H}} + P_i)^{-1}. \end{cases}$$

Note that the posterior mean \hat{c}_i coincides with the optimal c_i of (8) when all the other $\{c_j\}_{j \neq i}$ are assumed to be known.

3.4. Multi-task estimation using MCMC

In the above analysis we have assumed known hyper-parameters Σ and γ but in practice they have also to be estimated from data. One exception can be given by Γ that contains the interaction parameters γ_{ij} . If orthogonality among the tasks i, j is known, one can just set γ_{ij} to a large value. However, we now design a

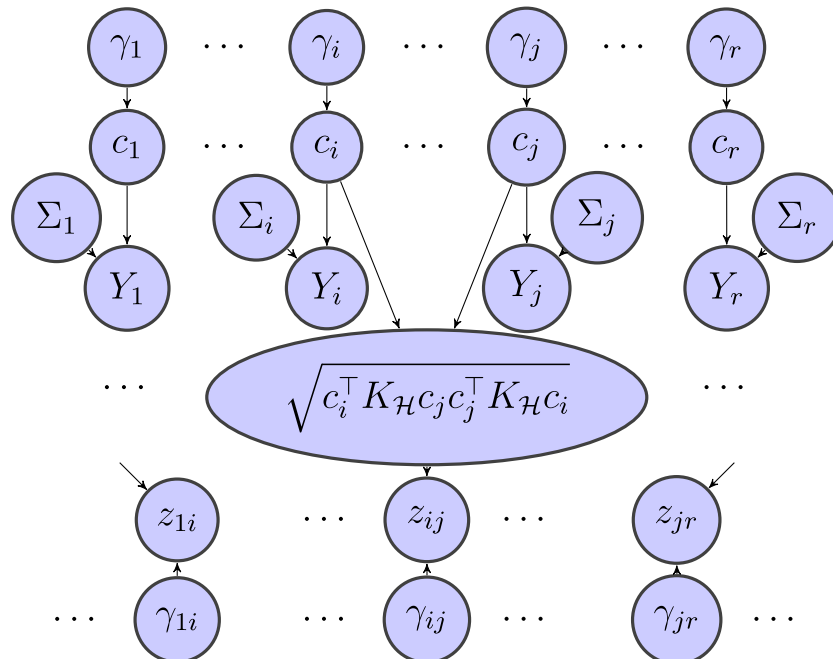


Fig. 1. Bayesian network describing the stochastic interpretation of problem (8). Each task i is represented by the random vector c_i which contains some expansion coefficients. Such a vector has to be reconstructed from the measurements Y_i that contain noisy and direct samples of the function f_i . Furthermore, each couple of task i, j interacts through the virtual measurements z_{ij} and the hyper-parameters γ_{ij} . When z_{ij} is zero, large values of γ_{ij} imply that the tasks i and j are in practice orthogonal.

scheme that can also estimate such parameters so that one has also to learn from data if orthogonality among some tasks really holds. For this purpose, we resort to the Markov Chain Monte Carlo paradigm. The procedure consists of two steps: first, we simulate the posterior $p(c, \gamma, \Gamma, \Sigma | Y, Z = 0) := \pi(c, \gamma, \Gamma, \Sigma)$, building a Markov chain whose invariant distribution is $\pi(c, \gamma, \Gamma, \Sigma)$; then we use N values of each c_i sampled from such a chain to approximate the posterior mean via Monte Carlo integration.

The first step is conveniently implemented via Gibbs sampling [54]. The rationale is to sequentially draw samples from the conditional distributions entering $\pi(c, \gamma, \Gamma, \Sigma)$. According to the decomposition of the joint distribution (10) given by the Bayesian network in Figure 1, these are.

- $\pi(c_i | c_{j \neq i}, \gamma_i, \Gamma, \Sigma_i)$ for each $i = 1, \dots, r$;
- $\pi(\gamma_i | c_i)$ for $i = 1, \dots, r$;
- $\pi(\sigma_{ki}^2 | c_i)$ for $i = 1, \dots, r$ and $k = 1, \dots, n$;
- $\pi(\gamma_{ij} | c_i, c_j)$ for $i = 1, \dots, r$ and $j > i$.

As seen in (12), $\pi(c_i | c_{j \neq i}, \gamma_i, \Gamma, \Sigma_i)$ is Gaussian and well defined. To deal with γ_i, γ_{ij} and σ_{ki}^2 for $i = 1, \dots, r, j > i$ and $k = 1, \dots, n$, properties of conjugate distributions can be leveraged as well. Assume that all of these hyper-parameters are endowed with an uninformative Gamma distribution over the positive real axis. Hence, denoting with $\text{Gamma}(a, b)$ a Gamma random variable with mean a/b , we get

$$\gamma_i | c_i \sim \text{Gamma}\left(\frac{n}{2}, \frac{c_i^\top K_{\mathcal{H}} c_i}{2}\right), \quad (13)$$

$$\sigma_{ki}^2 | c_i \sim \text{Gamma}\left(\frac{1}{2}, \frac{(y_{ki} - [K_{\mathcal{H}} c_i]_k)^2}{2}\right), \quad (14)$$

$$\gamma_{ij} | c_i, c_j \sim \text{Gamma}\left(\frac{1}{2}, \frac{c_i^\top K_{\mathcal{H}} c_j c_j^\top K_{\mathcal{H}} c_i}{2}\right). \quad (15)$$

The overall Gibbs sampling scheme is summarized in Algorithm 1.

Algorithm 1: Input: the number r , data-sets \mathcal{X} and \mathcal{Y} (also expressed in vectors $\{Y_i\}_{i=1}^r$) and the number M of MCMC iterations. Output: a stochastic simulator of the posterior $\pi(c, \Sigma, \gamma, \Gamma)$.

```

for  $m = 1, \dots, M$  do
  for  $i = 1, \dots, r$  do
    if  $m = 1$  then
      set  $c_i(1) = K_{\mathcal{H}}^{-1} Y_i$ ;
    else
      sample  $c_i(m)$  from (12);
    end if
    sample  $\gamma_i(m)$  from (13);
    sample  $\sigma_{ki}^2(m)$  from (14),  $k = 1, \dots, n$ ;
    build  $\Sigma_i(m)$ ;
  end for
  for  $i = 1, \dots, r$  and  $j > i$  do
    sample  $\gamma_{ij}(m)$  as in (15);
  end for
end for

```

All full conditionals admit a well-defined density on an open connected (sub) set of an Euclidean space. Then, it follows that the Markov chain is irreducible, i.e., that the support of the joint distribution can be entirely explored (see, e.g., [53][Chapter IV], [55,56]). Irreducibility is a sufficient condition for the Law of Large Numbers to hold [57], thus legitimating the use of Monte Carlo integration. This allows also to compute the estimates for the hyper-parameters γ_{ij} . The resulting values can be deployed to detect whether the orthogonality constraints are present. Eventually, this leads to a decision rule to discard the ones that are not active, e.g., whose γ_{ij} estimates are smaller than a certain threshold.

3.5. Bayesian network with reduced complexity

The Bayesian network displayed in Fig. 1 considers distinct all the $r(r-1)/2$ values of γ_{ij} . Hence, it aims to estimate all the inter-

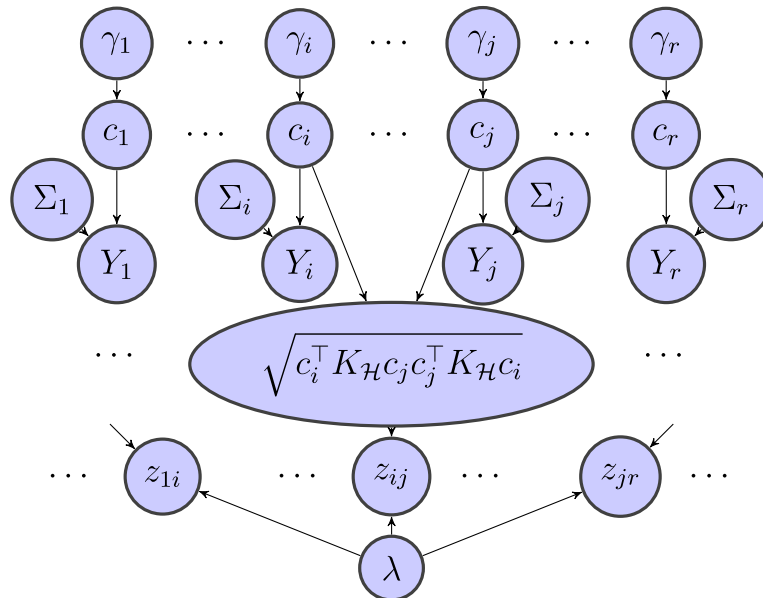


Fig. 2. Bayesian network obtained assuming $\gamma_{ij} = \lambda$ for any $i = 1, \dots, r$ and $j > i$. This model describes a situation where one postulates that all the tasks are orthogonal and then wants to learn from data if this assumption really holds true. When large estimated values of λ are returned by the MCMC schemes, the orthogonality assumption turns out to be confirmed.

actions among all the possible couple of tasks. This model can be too sophisticated, leading to estimates affected by large variance. Hence, it can be convenient to reduce its complexity. This can be achieved in several ways. For instance, some of the γ_{ij} could be fixed to zero, hence including the information that tasks i and j are known to be non orthogonal. One could also select a subset of tasks which are instead known to be orthogonal and fix all the related γ_{ij} to a common large value. Another significant situation arises when all the tasks are postulated to be orthogonal and one wants to learn from data if this assumption is true. This important model is described through the Bayesian network in Fig. 2 where a single value $\lambda = \gamma_{ij}$ is assigned to all the tasks couples. Thus, (15) becomes

$$\lambda | \mathbf{c} \sim \text{Gamma} \left(\frac{r(r-1)}{4}, \frac{\sum_{i=1}^r \sum_{j>i}^r c_i^T K_{\mathcal{H}} c_j c_j^T K_{\mathcal{H}} c_i}{2} \right). \quad (16)$$

This model version will be tested through the numerical experiments reported in Section 5.

4. Theoretical analysis on the infinite-dimensional orthogonal multi-task problem

In this section we investigate some theoretical issues connected with orthogonal multi-task learning. We are given r deterministic functions f_i belonging to the RKHS \mathcal{H} . They are also assumed orthogonal in the Hilbert space \mathcal{H}^r and, to simplify notation, of unit norm.

Given any $g \in \mathcal{H}^r$, we use \mathbb{K}_g to indicate the $r \times r$ Gram matrix whose (i, j) -entry is

$$[\mathbb{K}_g]_{ij} = \langle g_i, g_j \rangle_{\mathcal{H}}.$$

Under the stated orthogonality assumptions, it comes that \mathbb{K}_f is the $r \times r$ identity matrix.

The functions f_i have to be reconstructed starting from the measurements described through (1). Assume that all the data-set sizes n_i are equal to n for sake of simplicity and consider the following two-step procedure. At the first step, r decoupled estimators are exploited to obtain r function estimates. Support vector machines or regularization networks could be used. Using the latter, one has

$$\hat{f}_i^n = \arg \min_{f_i \in \mathcal{H}} \sum_{k=1}^n \frac{(y_{ki} - f_i(x_{ki}))^2}{\sigma_{ki}^2} + \gamma_i^n \|f_i\|_{\mathcal{H}}^2,$$

where we have also stressed the dependence of the regularization parameter and of the task estimates on the data-set size n .

At the second step, we interpret $\{\hat{f}_i^n\}_{i=1}^r$ as the r components of the infinite-dimensional data $\mathbf{y}^n \in \mathcal{H}^r$. Such data are exploited to obtain r orthogonal task estimates by solving the following infinite-dimensional orthogonal multi-task learning problem

$$\hat{\mathbf{f}}^n = \arg \min_{\mathbf{f} \in \mathcal{D} \subset \mathcal{H}^r} \|\mathbf{f} - \mathbf{y}^n\|_{\mathcal{H}^r}, \quad (17)$$

where

$$\mathcal{D} = \{\mathbf{f} \in \mathcal{H}^r \text{ s.t. } \mathbb{K}_{\mathbf{f}} = I_r\}. \quad (18)$$

The theoretical issues that are addressed in the remainder of this section are the following:

(Q1): Existence of $\hat{\mathbf{f}}^n$.

(Q2): Convergence of optimizer $\hat{\mathbf{f}}^n$ as $n \rightarrow +\infty$.

(Q1) is not trivial since (17) is a non-convex infinite-dimensional optimization problem, being subject to orthogonality constraints. As regards (Q2), it is intimately linked to the consistency properties of the decoupled estimators employed during the first step. In the case of the regularization networks, in [58] one can find rules for γ_i^n that ensure convergence of \hat{f}_i^n to f_i (e.g., in the RKHS norm) over the regions where the x_{ki} become dense. If holes are present in the probability density function of the x_{ki} , i.e., regions that are never sampled like those introduced in our numerical experiments, the decoupled estimators will return an extension coherent with the kernel of \mathcal{H} . In many applications, \mathcal{H}^r leads to inner-products among the tasks defined by the classical Lebesgue spaces. Convergence in RKHS norm implies convergence also in those spaces [4], hence one can assume that \mathbf{y}^n converges to an element \mathbf{y} in \mathcal{H}^r . Under this circumstance, questions (Q1) and (Q2) are then addressed in the following theorem.

Theorem 1. Problem (17) always admits at least one solution, i.e., the orthogonal multi-task estimates exist.

Furthermore, if \mathbf{y}^n converges to \mathbf{y} in \mathcal{H}^r , letting

$$\hat{\mathbf{f}} = \arg \min_{\mathbf{f} \in \mathcal{D} \subset \mathcal{H}^r} \|\mathbf{f} - \mathbf{y}\|_{\mathcal{H}^r}, \quad (19)$$

one has

$$\lim_{n \rightarrow \infty} \|\hat{\mathbf{f}} - \hat{\mathbf{f}}^n\|_{\mathcal{H}^r} = 0. \quad (20)$$

Proof: We first show (Q2) by assuming that the multi-task estimates $\hat{\mathbf{f}}^n$ in (17) exist. One easily has

$$\|\mathbf{f} - \mathbf{y}\|_{\mathcal{H}^r} \leq \|\mathbf{f} - \mathbf{y}^n\|_{\mathcal{H}^r} + \|\mathbf{y} - \mathbf{y}^n\|_{\mathcal{H}^r}$$

and

$$\|\mathbf{f} - \mathbf{y}^n\|_{\mathcal{H}^r} \leq \|\mathbf{f} - \hat{\mathbf{f}}^n\|_{\mathcal{H}^r} + \|\hat{\mathbf{f}}^n - \mathbf{y}^n\|_{\mathcal{H}^r},$$

that implies

$$\max_{\mathbf{f} \in \mathcal{H}^r} \|\|\mathbf{f} - \mathbf{y}^n\|_{\mathcal{H}^r} - \|\mathbf{f} - \mathbf{y}\|_{\mathcal{H}^r}\| \leq \|\mathbf{y} - \mathbf{y}^n\|_{\mathcal{H}^r}.$$

Since, by assumption, \mathbf{y}^n converges to \mathbf{y} in \mathcal{H}^r , the above inequality shows that the objective $\|\mathbf{f} - \mathbf{y}^n\|_{\mathcal{H}^r}$ in (17) converges uniformly to the objective $\|\mathbf{f} - \mathbf{y}\|_{\mathcal{H}^r}$ in (19). It is well known that this also implies convergence of the optimizers (see, e.g., [59][Section 7.E] or [60][Section 4]), so that $\hat{\mathbf{f}}^n$ indeed converges to $\hat{\mathbf{f}}$ in \mathcal{H}^r . The remainder of this section is devoted to proving (Q1), that is to showing that the optimizer in (17) is well defined.

Our first step is to consider a modified version of the problem (19) by enlarging the optimization domain. Define

$$E = \{\mathbf{f} \in \mathcal{H}^r \text{ s.t. } \mathbb{K}_{\mathbf{f}} \leq I_r\} \quad (21)$$

where, according to the Loewner order, given two symmetric matrices M, N of the same size, $M \leq N$ means that $N - M$ is positive semi-definite. Then, let us consider

$$\begin{aligned} \check{\mathbf{f}} &= \arg \min_{\mathbf{f} \in E \subset \mathcal{H}^r} \|\mathbf{f} - \mathbf{y}\|_{\mathcal{H}^r}, \text{ which is} & (22a) \\ &= \arg \max_{\mathbf{f} \in E \subset \mathcal{H}^r} \langle \mathbf{f}, \mathbf{y} \rangle_{\mathcal{H}^r} \end{aligned} \quad (22b)$$

because all functions f_i have unitary norm by hypothesis.

Clearly, the objective $\langle \mathbf{f}, \mathbf{y} \rangle_{\mathcal{H}^r}$ in (22b) is weakly continuous ([61], Chapter 12) and convex. If the optimization domain E is weakly compact, it follows that a solution $\check{\mathbf{f}}$ yielding the maximum exists ([62], Chapter 2, Theorem (W*)) and can be found at the boundary of E , which is defined as

$$bE = \{\mathbf{f} \in \mathcal{H}^r \text{ s.t. } \mathbb{K}_{\mathbf{f}} \leq I_r \text{ and exists s.t. } \lambda_i(\mathbb{K}_{\mathbf{f}}) = 1\}, \quad (23)$$

where $\lambda_i(\mathbb{K}_f)$ denotes the i -th eigenvalue of \mathbb{K}_f . Hence, we prove in the next lines that E is weakly compact.

First, note that

$$\|f\|_{\mathcal{H}^r}^2 \leq r \quad \forall f \in E$$

since

$$\|f\|_{\mathcal{H}}^2 = [\mathbb{K}_f]_{ii} \leq 1 \quad \forall f \in E.$$

So, the set E is bounded in \mathcal{H}^r . This implies that, given any sequence contained in E , we can extract from it a subsequence $\{h^{(i)}\}$ that converges weakly to a function $h \in \mathcal{H}^r$ ([62], Theorem 2.C). The crucial point is to show that one has also $h \in E$. By Mazur's Lemma ([63], Section V.2, Theorem 2), the limit function h corresponds to the strong limit of particular convex combinations of functions taken from $\{h^{(i)}\}$. In other words, there exists a subsequence $\{\bar{h}^{(i)}\}$ built with convex combinations of the $\{h^{(i)}\}$ such that

$$\lim_{i \rightarrow \infty} \|h - \bar{h}^{(i)}\|_{\mathcal{H}^r} = 0. \quad (24)$$

Now, let $f = h^{(k)}$ and $g = h^{(j)}$ (or, more generally, one can let f and g be two generic vectors in E). Such a couple contains the $2r$ functions $\{f_i, g_i\}_{i=1}^r$. Applying to such a set the Gram-Schmidt procedure, we obtain the functions $\{\eta_i\}_{i=1}^n$ orthonormal in \mathcal{H} . Using η to denote the function that embeds the $\{\eta_i\}$, we can write

$$f = M\eta, \quad g = N\eta$$

that means that f_i and g_i are the linear combinations of the $\{\eta_i\}$ with coefficients contained in the i -th row of the $r \times n$ matrices M and N , respectively. One thus has

$$\mathbb{K}_f = MM^\top, \quad \mathbb{K}_g = NN^\top.$$

Consider any convex combination of f and g , i.e. the functions $af + (1-a)g$ with $a \in [0, 1]$. Then, one has

$$\begin{aligned} \mathbb{K}_{af+(1-a)g} &= \mathbb{K}_{aM\eta+(1-a)N\eta} \\ &= (aM + (1-a)N)(aM + (1-a)N)^\top \\ &= aMM^\top + (1-a)NN^\top - a(1-a)(M-N)(M-N)^\top \\ &\leq aMM^\top + (1-a)NN^\top = a\mathbb{K}_f + (1-a)\mathbb{K}_g \\ &\leq aI_r + (1-a)I_r = I_r \end{aligned}$$

and this shows that

$$\bar{h}^{(i)} \in E \quad \text{for all } i. \quad (25)$$

Now we prove that $h \in E$. Let $\{f^{(i)}\}$ and $\{g^{(i)}\}$ be sequences in \mathcal{H}^r convergent to f and g , respectively. For any couple of functions $f_j^{(i)}$ and $g_k^{(i)}$, representing the j -th and k -th component of $f^{(i)}$ and $g^{(i)}$, respectively, one has

$$\begin{aligned} \langle f_j^{(i)}, g_k^{(i)} \rangle_{\mathcal{H}} &= \langle f_j^{(i)} - f_j + f_j, g_k^{(i)} - g_k + g_k \rangle_{\mathcal{H}} \\ &= \langle f_j^{(i)} - f_j, g_k^{(i)} - g_k \rangle_{\mathcal{H}} + \langle f_j^{(i)} - f_j, g_k \rangle_{\mathcal{H}} \\ &\quad + \langle f_j, g_k^{(i)} - g_k \rangle_{\mathcal{H}} \leq \|f_j^{(i)} - f_j\|_{\mathcal{H}} \|g_k^{(i)} - g_k\|_{\mathcal{H}} + \langle f_j^{(i)} - f_j, g_k \rangle_{\mathcal{H}} + \langle f_j, g_k^{(i)} - g_k \rangle_{\mathcal{H}} \end{aligned}$$

and this shows that $\langle f_j^{(i)}, g_k^{(i)} \rangle_{\mathcal{H}}$ converges to $\langle f_j, g_k \rangle_{\mathcal{H}}$ as i grows to $+\infty$. This fact, combined with (24), implies that

$$\mathbb{K}_{\bar{h}^{(i)}} \rightarrow \mathbb{K}_h \quad \text{for } i \rightarrow +\infty$$

where the symbol \rightarrow denotes convergence under any matrix norm. Using the fact that the entries and eigenvalues of the Gram matrix depend continuously on the $\bar{h}^{(i)}$, and recalling also (25), we obtain

$$\mathbb{K}_{\bar{h}^{(i)}} \leq I_r \quad \forall i \Rightarrow \mathbb{K}_h \leq I_r.$$

This shows that

$$h \in E$$

and proves the (desired) weak compactness of E . The consequence is that the solution of (22) exists and, since the objective is convex, it belongs to the boundary bE of E . So, there is a maximizer \check{f} of (22b) such that $\mathbb{K}_{\check{f}} \leq I_r$ and $\max_i \lambda_i(\mathbb{K}_{\check{f}}) = 1$. Now, we want to prove that another maximizer \hat{f} exists and satisfies $\mathbb{K}_{\hat{f}} = I_r$.

Given any set of functions $\{f_i\}_{i=1}^r$ in \mathcal{H} , embedded in $f \in \mathcal{H}^r$, they can be expressed as linear combination of orthonormal functions $\{\eta_i\}_{i=1}^r$, contained in η , with coefficients defined by a symmetric and positive semi-definite matrix.¹ Specifically, considering our optimizer \check{f} in place of the generic f , we can write

$$\check{f} = M\eta, \quad M = M^\top \geq 0.$$

Consider the SVD $M = VD V^\top$, where D is diagonal with (i, i) -entry given by d_i , and define $\Delta = \eta - \check{f} = (I_r - M)\eta$. Then, we have

$$\begin{aligned} \mathbb{K}_{\check{f}+a\Delta} &= \mathbb{K}_{(M+a(I_r-M))\eta} \\ &= (M + a(I_r - M))(M + a(I_r - M))^\top \\ &= (aI_r + (1-a)M)(aI_r + (1-a)M)^\top \\ &= (aI_r + (1-a)M)^2 \\ &= V \text{diag}\left(\left\{a + (1-a)d_i\right\}_{i=1}^r\right) V^\top. \end{aligned}$$

So, the r eigenvalues of $\mathbb{K}_{\check{f}+a\Delta}$ are $(a + (1-a)d_i)^2$ and one has

- the function $\check{f} + a\Delta$ belongs to bE for any $a \in [-1, 1]$. In fact, since $0 \leq d_i \leq 1$ and there exists j s.t. $d_j = 1$, one has

$$(a + (1-a)d_i)^2 \leq 1 \quad \forall a \in [-1, 1]$$

and also

$$(a + (1-a)d_j)^2 = (a + (1-a))^2 = 1 \quad \forall a \in [-1, 1].$$

- the function $\check{f} + \Delta$ not only belongs to bE but also to our original optimization domain D . In fact, it satisfies the constraint $\mathbb{K}_{\check{f}+\Delta} = I_r$ since, by setting $a = 1$, one obtains

$$(a + (1-a)d_i)^2 = (1 + (1-1)d_i)^2 = 1 \quad i = 1, \dots, r.$$

Now, we will show that the (possibly not unique) solution \hat{f} of (19) is indeed given by $\check{f} + \Delta$.

Consider the objective in (22b) and, to simplify the notation, denote it by $J(f) = \langle f, y \rangle_{\mathcal{H}^r}$. Since J is convex and \check{f} maximizes J over E , one obtains

$$J(\check{f}) \geq J(\check{f} + \Delta), \quad J(\check{f}) \geq J(\check{f} - \Delta).$$

But J is convex and this implies that

$$J(a(\check{f} + \Delta) + (1-a)(\check{f} - \Delta)) \leq aJ(\check{f} + \Delta) + (1-a)J(\check{f} - \Delta)$$

for any $a \in [0, 1]$. With $a = 1/2$ we obtain

$$J(\check{f}) \leq \frac{1}{2}(J(\check{f} + \Delta) + J(\check{f} - \Delta))$$

¹ In fact, by Gram-Schmidt one has $f = A\zeta$ where ζ contains r orthonormal functions in \mathcal{H} (if f does not contain r independent functions, we just add some orthonormal functions to those returned by Gram-Schmidt to obtain r orthonormal functions and form $\zeta \in \mathcal{H}^r$). Consider the SVD $A = UDV^\top = UDU^\top UV^\top$. Then, define $M = UDU^\top, \eta = UV^\top \zeta$ and note that $\mathbb{K}_\eta = UV^\top VU^\top = I_r$.

and then one must have

$$J(\tilde{f}) = J(\tilde{f} + \Delta) = J(\tilde{f} - \Delta).$$

Hence, the function $\tilde{f} + \Delta$ not only belongs to D but is also a maximizer of J . This completes the proof.

5. Numerical experiments

We now test the effectiveness of the computational scheme presented in Section 3. In the following experiments, we compare the performance of the proposed orthogonality-constrained approach with the standard single-task estimation recalled in Section 3.1. The Bayesian model depicted in Fig. 2 will be adopted. Hence, tasks can be all mutually orthogonal but this information has to be corroborated from data.

We consider a population of $r = 20$ tasks. For each of them, $n = 100$ noisy measurements are collected according to model (1). Input locations are independent realizations from the uniform distribution over $\mathcal{X} = [0, 1]$ and the noise variance is $\sigma^2 = 0.04$. In any experiment, the functions belong to the Hilbert space $\mathcal{H} = \mathcal{L}^2$ with inner product $\langle f_i, f_j \rangle_{\mathcal{H}} = \int_0^1 f_i(x)f_j(x)dx$. The RKHS \mathcal{H} carrying information about function smoothness is the Sobolev space induced by the spline kernel

$$\mathcal{K}(x_a, x_b) = \min(x_a, x_b) \quad \text{with } x_a, x_b \in [0, 1]. \tag{26}$$

Such a space is known to be contained in \mathcal{L}^2 , so the problem is well posed. In this set-up, regularizer $K_{\mathcal{H}}$ entering (8) is the kernel matrix associated to (26), while $K_{\mathcal{H}}$ is derived from (7) as the \mathcal{L}^2 -inner product of kernel sections in \mathcal{H} . These are composed by ramp and constant functions, so the (a, b) element of matrix $K_{\mathcal{H}}$ is

$$\begin{aligned} [K_{\mathcal{H}}]_{a,b} &= \int_0^1 \mathcal{K}(x_a, x)\mathcal{K}(x_b, x)dx \\ &= \frac{x_b^3}{3} + x_a^2(x_b - x_a) + \frac{x_b}{2}(x_b - x_a)^2 + x_ax_b(1 - x_b) \end{aligned}$$

for $a, b = 1, \dots, n$. As a performance metric, we consider for the i -th task

$$Fit = 100\% \left(1 - \frac{\|f_i - \hat{f}_i\|_2}{\|f_i\|_2} \right).$$

In practice, the \mathcal{L}^2 norm will be numerically computed by taking the Euclidean norm on the vectors containing the pointwise function evaluations over a grid of 1000 equispaced samples of \mathcal{X} .

5.1. First test: non-orthogonal functions

In the first experiment, the functions to be estimated are

$$f_i(x) = x^i \quad i = 1, \dots, r. \tag{27}$$

Since $\langle f_i, f_j \rangle_{\mathcal{H}} \neq 0$ for all $i \neq j$, the orthogonality constraint included in the multi-task approach could in principle undermine the estimation performance with respect to the single-task case. On the contrary, the results of this test show that the proposed method is capable to disable the orthogonality constraint, hence yielding a fit score comparable to that returned by the single-task approach. Fig. 3 reports the boxplots of the average fits over all 20 tasks for both approaches on a sample run. Being the mean fits equal to 77.59 and 76.62 for the single- and multi-task cases respectively, one can appreciate the flexibility of the proposed approach. To further visualize the results, Fig. 4 displays the estimation performance for task of indexes $i = 1, 3, 4, 5, 6, 7, 8, 10$.

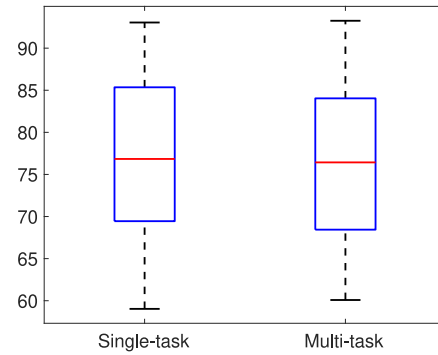


Fig. 3. Fit scores for single- and multi-task estimates considering power functions (27). A rich data-set is available for any task.

5.2. Second test: orthogonal functions

In the second experiment, the tasks are orthogonal, defined by

$$f_i(x) = \sin(2\pi xi) \quad i = 1, \dots, r. \tag{28}$$

The results on a single run are presented in Fig. 5 and Fig. 6, displaying the boxplot of fit scores over the 20 tasks and the fit performance, respectively. In the presented run, the mean scores are 78.87 and 79.83 for the single- and multi-task approach.

We can note that, even if the orthogonality constraint is active, it does not yield an impressive improvement with respect to the single-task set-up. This is due to the fact that a good amount of noisy and direct measurements, well distributed over all the unit interval, are available for any task. To point out advantages of the orthogonality constraints, the following test introduces a more complex scenario.

5.3. Third test: orthogonal functions with missing measurements

We now consider the same functions presented in (28). The measurements model is the same but, after generating the 100 input locations, we assume that measurements corresponding to a randomly placed window of 30 adjacent samples over \mathcal{X} are missing for tasks $i = 1, 2, 3, 6, 7, 8, 9, 10$. On these, while single-task learning is quite challenging, we will see that the multi-task approach leverages the additional information about orthogonality and copes better with the incomplete data-set. Fig. 7 reports the boxplots for the fit scores over the most challenging estimation problems, i.e. those related to tasks $i = 1, 2, 3, 6, 7, 8, 9, 10$. Next, Fig. 8 presents a sample performance corresponding to one realization of noises and missing data. On this run, the average fit scores are 61.13 and 75.94 for the single- and multi-task approach, respectively.

Within this framework, we finally assess the overall performance by running a Monte Carlo experiment of 100 runs. At any run, new independent noise realizations are generated. Fig. 9 shows the boxplot of the average fit scores. The orthogonality-constrained approach clearly outperforms single-task learning: the median improvement is of 16%, and one can also see a significative reduction of the estimator's variance.

5.4. Fourth test: orthogonal functions estimation, comparison with deterministic optimization

We now aim at estimating $r = 8$ unknown orthogonal functions defined in (28) from $n = 100$ direct and noisy measurements. The goal of this test is to compare the performance of the proposed

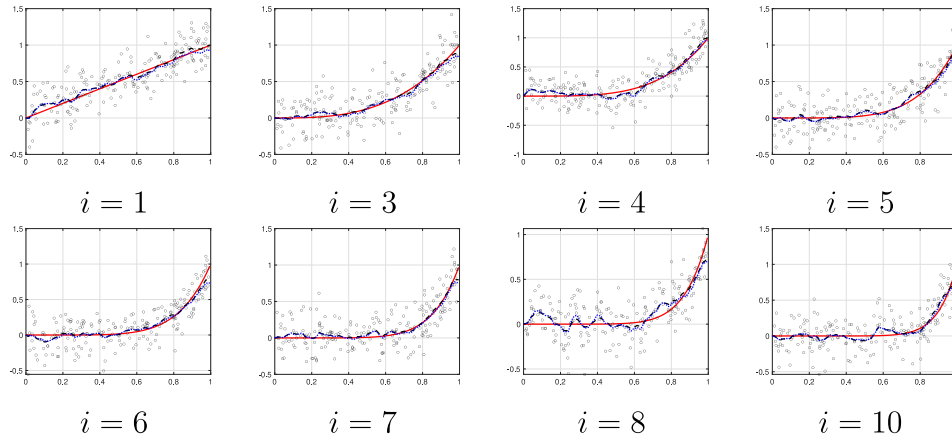


Fig. 4. Sample fitting performance in the set-up presented in Section 5.1. LEGEND: Black circles = data points; Solid red = true function; Dashed black = single-task estimate; Dash-dotted blue = multi-task estimate.

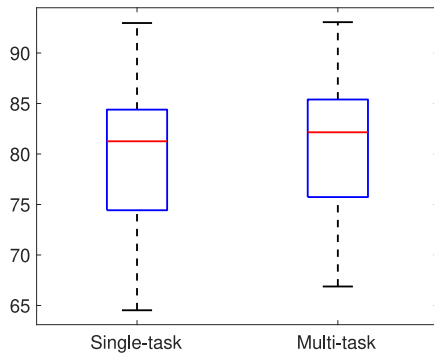


Fig. 5. Fit scores for single- and multi-task estimates for sinusoidal functions (28). A rich data-set is available for any task.

MCMC-based approach with the one yielded by a deterministic optimization routine.

We take the problem as stated in (8), again setting $\gamma_{ij} = \lambda$ for all $i = 1, \dots, r$ and $j > i$. We run the proposed multi-task learning routine based on MCMC, and fix the values for $\{\gamma_i\}_{i=1}^r$ and λ that were estimated in the output of the scheme. Next, we solve (8) using

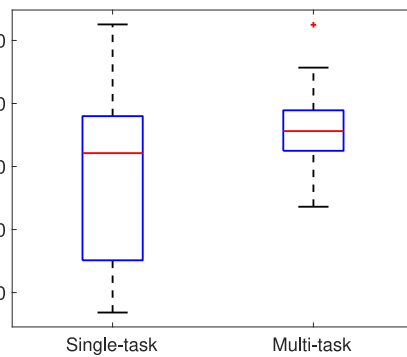


Fig. 7. Fit scores for single- and multi-task estimates considering sinusoidal functions (28) with incomplete data-sets. A single run is considered and the fit score is measured only on tasks 1,2,3,4,6,7,8,9,10, i.e. on the ones involving missing data.

those hyper-parameters values, deploying `fminunc` in Matlab as a deterministic optimization routine. The overall optimization routine is initialized at c_i equal to the zero vector for each $i = 1, \dots, r$. Fig. 10 displays a sample performance corresponding to one realization of noises, merging the fit scores for all tasks.

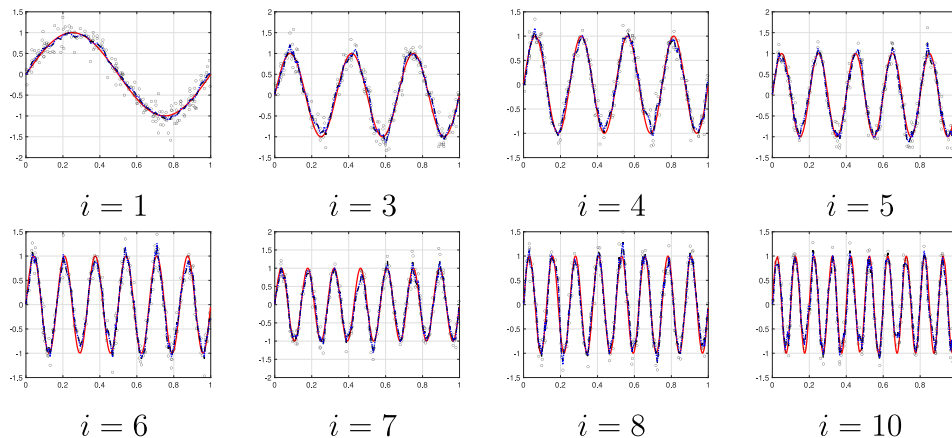


Fig. 6. Sample fitting performance in the set-up presented in Section 5.2. LEGEND: Black circles = data points; Solid red = true function; Dashed black = single-task estimate; Dash-dotted blue = multi-task estimate.

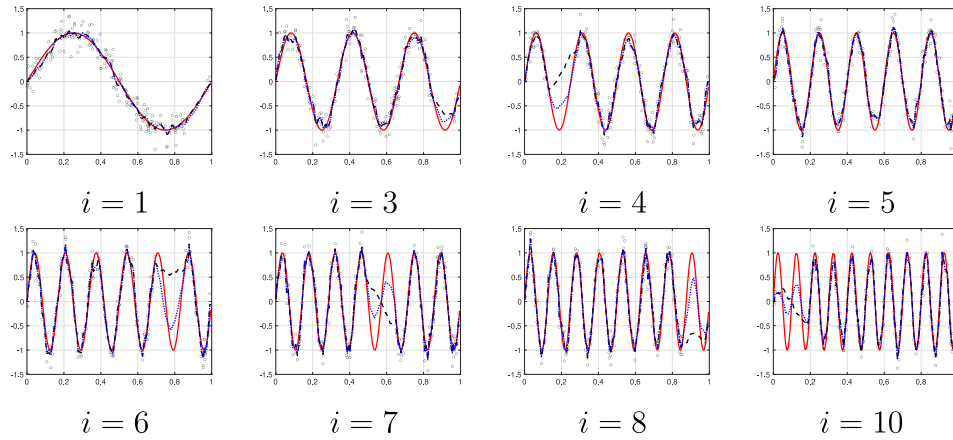


Fig. 8. Sample fitting performance in the set-up presented in Section 5.3. LEGEND: Black circles = data points; Solid red = true function; Dashed black = single-task estimate; Dash-dotted blue = multi-task estimate.

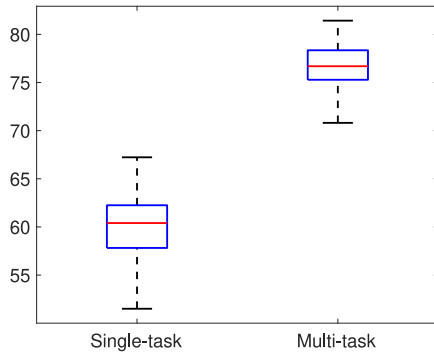


Fig. 9. Boxplot of the fits achieved after 100 Monte Carlo runs in the set-up presented in Section 5.3.

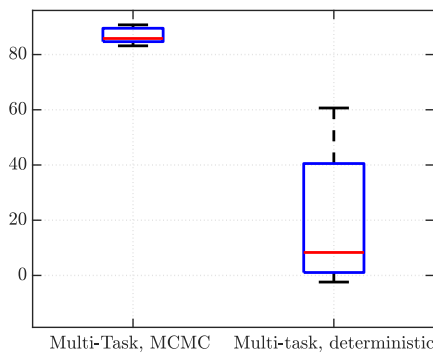


Fig. 10. Fit scores for multi-task estimates for sinusoidal functions (28) considering the proposed approach based on the Gibbs sampler versus a deterministic optimization scheme.

The MCMC-based approach clearly outperforms the one based on deterministic optimization: this is because the latter is not able to overcome the intrinsic non-convexity of the problem, and is thus sensitive to initialization. Moreover, it also becomes computationally prohibitive for large values of n and/or r , because optimization is taken over $[c_1^T, \dots, c_r^T] \in \mathbb{R}^{nr}$. This does not happen in the

proposed approach based on MCMC, because there each c_i is sampled separately. We also point out that our scheme automatically includes hyper-parameters estimation, which is another intricate problem that would further encumber deterministic optimization routines.

6. Conclusions

We have provided new algorithmic and theoretical results regarding orthogonal multi-task learning, that is the problem of reconstructing orthogonal functions in Hilbert spaces. Differently from standard multi-task problems, the main difficulty is its intrinsic non convexity related to the complex nonlinear interactions among the tasks.

To include orthogonality information, we have complemented the classical kernel-based estimators for nonparametric function estimation with special regularization terms that also enjoy a Bayesian interpretation. In fact, the orthogonality constraints derive from particular a priori probability density functions. When data become available, the posterior of any task (conditional on all the others) remains Gaussian. This allows the design of a new stochastic simulation scheme based on Gibbs sampling that overcomes non convexity by returning the joint posterior of all the unknown functions in sampled form. Numerical results show the goodness of the new approach also revealing its potentiality to detect from data if tasks orthogonality really holds.

CRedit authorship contribution statement

Anna Scampicchio: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation. **Mauro Bisiacco:** Conceptualization, Writing - review & editing, Supervision. **Gianluigi Pilonetto:** Conceptualization, Methodology, Software, Supervision, Project administration.

Data availability

No data was used for the research described in the article.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was financially supported by the University of Padova and ETH Zürich.

References

- [1] T.J. Hastie, R.J. Tibshirani, J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, Canada, 2001.
- [2] B. Schölkopf, A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, (Adaptive Computation and Machine Learning), MIT Press, 2001.
- [3] N. Aronszajn, Theory of reproducing kernels, *Transactions of the American Mathematical Society* 68 (1950) 337–404.
- [4] F. Cucker, S. Smale, On the mathematical foundations of learning, *Bulletin of the American mathematical society* 39 (2001) 1–49.
- [5] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, *Advances in Computational Mathematics* 13 (2000) 1–50.
- [6] R. Caruana, Multitask learning, *Machine Language* 28 (1) (1997) 41–75.
- [7] S. Thrun, L. Pratt, *Learning to learn*, Kluwer, 1997.
- [8] B. Bakker, T. Heskes, Task clustering and gating for Bayesian multitask learning, *Journal of Machine Learning Research* 4 (2003) 83–99.
- [9] Y. Zhang, Q. Yang, A survey on multi-task learning, *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1.
- [10] C. Micchelli, M. Pontil, On learning vector-valued functions, *Neural Computation* 17 (1) (2005) 177–204.
- [11] T. Evgeniou, C. Micchelli, M. Pontil, Learning multiple tasks with kernel methods, *Journal of Machine Learning Research* 6 (2005) 615–637.
- [12] M.A. Álvarez, L. Rosasco, N.D. Lawrence, *Kernels for Vector-Valued Functions: A Review*, Vol. 4, Now Publishers Inc., Hanover, MA, USA, 2012.
- [13] Z. Liu, B. Huang, Y. Cui, Y. Xu, B. Zhang, L. Zhu, Y. Wang, L. Jin, D. Wu, Multi-task deep learning with dynamic programming for embryo early development stage classification from time-lapse videos, *IEEE Access* 7 (2019) 122153–122163.
- [14] C. Karmakar, B. Saha, M. Palaniswami, S. Venkatesh, Multi-task transfer learning for in-hospital-death prediction of icu patients, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2016* (2016) 3321/3324.
- [15] C. Williams, S. Klanke, S. Vijayakumar, K. Chai, Multi-task Gaussian process learning of robot inverse dynamics, in: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (Eds.), *Advances in Neural Information Processing Systems*, Vol. 21, Curran Associates Inc, 2009.
- [16] W.H. Greene, *Econometric Analysis*, 5th Edition, Pearson Education, 2003.
- [17] D. Fourure, R. Emonet, E. Fromont, D. Muselet, N. Neverova, A. Treméau, C. Wolf, Multi-task, multi-domain learning: application to semantic segmentation and pose regression, *Neurocomputing* 251 (2017) 68–80.
- [18] J. Ghosn, Y. Bengio, Multi-task learning for stock selection, in: *Proceedings of the 9th International Conference on Neural Information Processing Systems, NIPS'96*, MIT Press, Cambridge, MA, USA, 1996.
- [19] H. Ling, Z. Wang, P. Li, Y. Shi, J. Chen, F. Zou, Improving person re-identification by multi-task learning, *Neurocomputing* 347 (2019) 109–118.
- [20] A. Gelman, J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Analytical Methods for Social Research, Cambridge University Press, 2006.
- [21] A.F. Zuur, E.N. Ieno, N.J. Walker, A.A. Saveliev, G.M. Smith, *Mixed Effects Modelling for Nested Data*, Springer New York, New York, NY, 2009, pp. 101–142.
- [22] *Nature Medicine* (2021) 94–105.
- [23] R. DerSimonian, N. Laird, Meta-analysis in clinical trials, *Controlled Clinical Trials* 7 (3) (1986) 177–188.
- [24] P. Zheng, R. Barber, R.J.D. Sorensen, C.J.L. Murray, A.Y. Aravkin, Trimmed constrained mixed effects models: Formulations and algorithms, *Journal of Computational and Graphical Statistics* 30 (3) (2021) 544–556.
- [25] G. Rosa, C. Padovani, D. Gianola, Robust linear mixed models with normal/independent distributions and bayesian mcmc implementation, *Biometrical Journal* 45 (5) (2003) 573–590.
- [26] Y.W. Teh, M.W. Seeger, M.I. Jordan, Semiparametric latent factor models, in: *AISTATS*, 2005.
- [27] G. Pillonetto, F. Dinuzzo, G. De Nicolao, Bayesian on-line multi-task learning of Gaussian processes, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (2) (2010) 193–205.
- [28] Q. Zhou, Q. Zhao, Flexible clustered multi-task learning by learning representative tasks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (2) (2016) 266–278.
- [29] A. Maurer, M. Pontil, B. Romera-Paredes, The benefit of multitask representation learning, *Journal of Machine Learning Research* 17 (1) (2016) 2853–2884.
- [30] J. Zhang, Y. Zhang, D. Ji, M. Liu, Multi-task and multi-view training for end-to-end relation extraction, *Neurocomputing* 364 (2019) 245–253.
- [31] J.C. Gower, G.B. Dijkstra, *Procrustes problems*, Vol. 30, Oxford University Press, Oxford, UK, 2004, of Oxford Statistical Science Series, URL:.
- [32] B. Tang, G. Sapiro, V. Caselles, Color image enhancement via chromaticity diffusion, *IEEE Transactions on Image Processing* 10 (5) (2001) 701–707.
- [33] V. Ozolins, R. Lai, R. Cafilisch, S. Osher, Compressed modes for variational problems in mathematics and physics, in: *Proceedings of the National Academy of Sciences of the U.S.A.*, Vol. 110, 2013, pp. 18368–73.
- [34] P.H. Schönemann, A generalized solution of the orthogonal procrustes problem, *Psychometrika* 31 (1966) 1–10.
- [35] A. Wu, R. Liu, Y. Han, L. Zhu, Y. Yang, Vector-decomposed disentanglement for domain-invariant object detection, in: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9322–9331. doi:10.1109/ICCV48922.2021.00921.
- [36] A. Wu, Y. Han, L. Zhu, Y. Yang, Instance-invariant domain adaptive object detection via progressive disentanglement, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (8) (2022) 4178–4193, <https://doi.org/10.1109/TPAMI.2021.3060446>.
- [37] M. Aharon, M. Elad, A. Bruckstein, K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, *IEEE Transactions on Signal Processing* 54 (11) (2006) 4311–4322.
- [38] P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, USA, 2007.
- [39] R. Lai, S. Osher, A splitting method for orthogonality constrained problems, *Journal of Scientific Computing* 58 (2) (2014) 431–449.
- [40] A. Edelman, T. Arias, S. Smith, The geometry of algorithms with orthogonality constraints, *SIAM Journal on Matrix Analysis and Applications* 20 (1998) 303–353.
- [41] P. Absil, R. Mahony, R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2008.
- [42] Z. Wen, W. Yin, A feasible method for optimization with orthogonality constraints, *Mathematical Programming* 142 (2013) 397–434.
- [43] H. Yuan, X. Gu, R. Lai, Z. Wen, Global optimization with orthogonality constraints via stochastic diffusion on manifold, *Journal of Scientific Computing* 80 (2) (2019) 1139–1170.
- [44] L. Eldén, H. Park, A procrustes problem on the stiefel manifold, *Numerische Mathematik* 82 (1999) 599–619.
- [45] A. Raftery, S. Lewis, *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall, 1996, Ch. Implementing MCMC.
- [46] A. Tikhonov, V. Arsenin, *Solutions of Ill-Posed Problems*, Winston/Wiley, Washington, D.C., 1977.
- [47] S. Berberian, *Introduction to Hilbert Space*, Oxford University Press, 1961.
- [48] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, Singapore, 1987.
- [49] G. Wahba, Y. Wang, Representer Theorem, *American Cancer Society* (2019) 1–11, <https://doi.org/10.1002/9781118445112.stat08200>.
- [50] A.Y. Aravkin, B.M. Bell, J.V. Burke, G. Pillonetto, The connection between bayesian estimation of a gaussian random field and RKHS, *IEEE Transactions on Neural Networks and Learning Systems* 26 (7) (2015) 1518–1524.
- [51] B.D.O. Anderson, J.B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, N.J., USA, 1979.
- [52] P. Magni, R. Bellazzi, G. De Nicolao, Bayesian function learning using MCMC methods, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (12) (1998) 1319–1331.
- [53] W. Gilks, S. Richardson, D. Spiegelhalter, *Markov chain Monte Carlo in Practice*, Chapman and Hall, London, 1996.
- [54] A.E. Gelfand, A.F.M. Smith, Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* 85 (410) (1990) 398–409.
- [55] K.S. Chan, Asymptotic behavior of the gibbs sampler, *Journal of the American Statistical Association* 88 (421) (1993) 320–326.
- [56] G. Roberts, A. Smith, Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms, *Stochastic Processes and their Applications* 49 (2) (1994) 207–216.
- [57] L. Tierney, Markov Chains for Exploring Posterior Distributions, *The Annals of Statistics* 22 (4) (1994) 1701–1728.
- [58] S. Smale, D. Zhou, Learning theory estimates via integral operators and their approximations, *Constructive Approximation* 26 (2007) 153–172.
- [59] R. Rockafellar, R.J.-B. Wets, *Variational Analysis*, Springer Verlag, Heidelberg, Berlin, New York, 1998.
- [60] H. Attouch, R.J.-B. Wets, Epigraphical analysis, *Annales de l'I.H.P. Analyse non linéaire* 5 (6) (1989) 73–100.
- [61] P. Lax, *Functional Analysis*, Wiley, 2014.
- [62] E. Zeidler, *Applied Functional Analysis*, Springer, 1995.
- [63] K. Yosida, *Functional Analysis*, Springer, Berlin Heidelberg, 1995.



Anna Scampicchio was born in 1993. She received in 2015 the Bachelor degree in Information Engineering and in 2017 the Masters degree in Automation Engineering, both cum laude, from the University of Padova. In 2017 she was awarded with the Roberto Rocca scholarship for her career during the Masters Degree. She held a visiting position at the Department of Applied Mathematics of University of Washington, Seattle, in 2019. In 2021 she received the Ph.D. in Information Engineering from the University of Padova and now she is a postdoctoral researcher at the Institute for Dynamic Systems and Control, ETH Zürich. Her research interests lie at the interplay among system identification, machine learning and control design.



Mauro Bisiacco received the Master Degree In Electronic Engineering in 1983 from the University of Padova (Italy). Since 1987 until 1990 he was Associate Professor of System Theory at the Engineering Faculty of the University of Udine (Italy). Since 1990 onward he is holding the same position at the University of Padova. His teaching activity includes many courses in the Automatic Control area (System Theory, Identification Theory, Automatic Control, Advanced Control Techniques, Systems and Models). His research activity is concerned with many aspects of System Theory. He authored/co-authored both more than 30 papers in

International Journals, and more than 30 contributions to International Conferences.



Gianluigi Pillonetto was born on January 21, 1975 in Montebelluna (TV), Italy. He received the Doctoral degree in Computer Science Engineering cum laude from the University of Padova in 1998 and the PhD degree in Bioengineering from the Polytechnic of Milan in 2002. In 2000 and 2002 he was visiting scholar and visiting scientist, respectively, at the Applied Physics Laboratory, University of Washington, Seattle. From 2002 to 2005 he was Research Associate at the Department of Information Engineering, University of Padova, becoming an Assistant Professor in 2005. He is currently a Full Professor of Control and Dynamic Systems at the Department of Information Engineering, University of Padova. His research interests are in the field of system identification, estimation and machine learning. He was Associate Editor of *Systems & Control Letters* and *IEEE Transactions on Automatic Control*. He currently serves as Associate Editor for *Automatica*. In 2003 he received the Paolo Durst award for the best Italian Ph.D. thesis in Bioengineering, he was the 2017 recipient of the *Automatica* Prize, assigned every three years for outstanding contributions to control theory by the International Federation of Automatic Control (IFAC) and *Automatica* (Elsevier), and he was Plenary Speaker at System Identification IFAC Symposium in 2018. He has been elevated to IEEE Fellow in 2020 for contributions to System Identification.