Equivalent between constrained optimal smoothing and Bayesian estimation

Laurence Grammont, Hassan Maatouk, Xavier Bay

Equivalence between constrained optimal smoothing and Bayesian estimation

Laurence Grammont[1], **Hassan Maatouk**[2] **and Xavier Bay**[3]

[1] *Université de Lyon, Institut Camille Jordan, UMR 5208, 42023 St-Étienne, France , e-mail:* [a]laurence.grammont@univ-st-etienne.fr

[2] *Department of Mathematics, CY Cergy Paris Université, Site du Parc, 95011 Cergy-Pontoise, France, e-mail:* [b]hassan.maatouk@cyu.fr

[3] *Mines Saint-Étienne, CNRS, UMR 6158 LIMOS, 42023 Saint-Étienne, France, e-mail:* [c]bay@emse.fr

Abstract: In this paper, we extend the correspondence between Bayesian estimation and optimal smoothing in a Reproducing Kernel Hilbert Space (RKHS) by adding convex constraints to the problem. Through a sequence of approximating Hilbertian subspaces and a discretized model, we prove that the *Maximum a posteriori* (MAP) of the posterior distribution is exactly the optimal constrained smoothing function in the RKHS. This paper can be read as a generalization of the paper [14], where it is proved that the optimal smoothing solution is the mean of the posterior distribution. Synthetic and real data studies confirm the correspondence established in this paper.

MSC2020 subject classifications: Primary 00X00, 00X00; secondary 00X00.
Keywords and phrases: Correspondence, inequality constraints, nonparametric regression, Reproducing Kernel Hilbert Space.

1. Introduction

Gaussian processes (GPs) are a powerful type of Bayesian model that is widely used in the machine learning community for nonparametric function estimation [26]. The flexibility of the covariance functions of GPs allows to incorporate prior assumptions, such as regularity, stationarity, sparsity, and constraints related to derivative information [10, 26]. Incorporating shape constraints into GPs like monotonicity, convexity and boundedness improves significantly the prediction accuracy and provides more realistic credible intervals [12, 21, 19, 17, 15, 33, 28, 16]. Recently, the authors in [31] conducted a survey of various methods for incorporating shape constraints into GPs.

In this paper, the finite-dimensional GP approximation originally proposed in [21] is considered. It is able to incorporate various shape constraints, such as monotonicity, boundedness, and convexity, in the entire domain. This approach has been considered in many recent papers [19, 18, 27, 33] and has been tested on real-world applications in different domains [7, 8, 9, 32]. The authors in

[18] have extended this approach to address multiple constraints. In the present paper, the asymptotic behavior of this approach has been investigated. The generalization of the Kimeldorf-Wahba correspondence between Bayesian estimation and optimal smoothing [14] has been established. We prove that the maximum a posteriori (MAP) estimate converges to the constrained optimal smooth function in the reproducing kernel Hilbert space (RKHS).

The paper is organized as follows: in Section 2, we briefly review GP regression. In Section 3, we present the framework of the correspondence. Section 4 is devoted to constructing the finite-dimensional GP approximation. In Section 5, the asymptotic analysis of the proposed method is investigated and the correspondence between Bayesian estimation and constrained optimal smoothing is established. Section 6 is devoted to numerical experiments, where the performance of the MAP estimate in terms of prediction accuracy is highlighted. Finally, in Section 7, a real data study is presented.

2. Gaussian process review

A GP namely Y is characterized by its mean function μ and covariance function k, i.e., $Y \sim \mathcal{GP}(\mu, k)$. It can be written as follows:

$$Y(\cdot) = \mu(\cdot) + U(\cdot),$$

where U is a zero-mean GP with covariance function k, i.e., $U \sim \mathcal{GP}(0, k)$,

$$k(\cdot,\cdot) = \mathrm{Cov}(U(\cdot), U(\cdot)) = \mathrm{E}[U(\cdot)U(\cdot)],$$

where Cov and E denote respectively the covariance and the expectation. The following regression problem is considered

$$y_i = u(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}}), \tag{1}$$

where u is an underlying function, $\{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n\}$ is a dataset of size n, with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ and ϵ_i is an additive and independent Gaussian noise with constant noise variance σ^2_{noise}. We denote by $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ the vector of data and by $\mathbb{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ the $n \times d$ matrix of designs, where d represents the dimension. A GP prior on the underlying function u is assumed. Conditionally on $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$, the conditional process is still a GP [14, 26]

$$\{U(\cdot) \mid U(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\} \sim \mathcal{GP}(\tilde{u}(\cdot), \tilde{k}(\cdot,\cdot)),$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top$ is a zero-mean Gaussian vector $\mathcal{N}(\mathbf{0}_n, \sigma^2_{\text{noise}}\mathbf{I}_n)$, with $\mathbf{I}_n$ the $n \times n$ identity matrix, $\mathbf{0}_n = [0, \ldots, 0]^\top$ the n-dimensional zero vector, and where $\tilde{u}$ and $\tilde{k}$ are respectively the conditional mean and covariance functions

$$\begin{aligned} \tilde{u}(\cdot) &= \mathrm{E}[U(\cdot) \mid U(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}] = k(\cdot, \mathbb{X})^\top \left(k(\mathbb{X}, \mathbb{X}) + \sigma^2_{\text{noise}}\mathbf{I}_n\right)^{-1} \boldsymbol{y}; \\ \tilde{k}(\cdot,\cdot) &= k(\cdot,\cdot) - k(\cdot, \mathbb{X})^\top (k(\mathbb{X}, \mathbb{X}) + \sigma^2_{\text{noise}}\mathbf{I}_n)^{-1} k(\cdot, \mathbb{X}). \end{aligned} \tag{2}$$

TABLE 1
Some popular covariance functions with their degree of smoothness [26] in one-dimensional case, i.e., $x, x' \in \mathbb{R}$.

Name	Expression	Class
Squared Exponential	$\exp\left(-\frac{(x-x')^2}{2\theta^2}\right)$	C^∞
Matérn $\nu = 5/2$	$\left(1+\frac{\sqrt{5}\lvert x-x'\rvert}{\theta}+\frac{5(x-x')^2}{3\theta^2}\right)\exp\left(-\frac{\sqrt{5}\lvert x-x'\rvert}{\theta}\right)$	C^2
Matérn $\nu = 3/2$	$\left(1+\frac{\sqrt{3}\lvert x-x'\rvert}{\theta}\right)\exp\left(-\frac{\sqrt{3}\lvert x-x'\rvert}{\theta}\right)$	C^1
Exponential	$\exp\left(-\frac{\lvert x-x'\rvert}{\theta}\right)$	C^0

Let us recall that $k(\mathbb{X}, \mathbb{X})$ is the covariance matrix of $U(\mathbb{X}) \in \mathbb{R}^n$ of dimension $n \times n$ and $k(\cdot, \mathbb{X})$ is the vector of covariance between $U(\cdot)$ and $U(\mathbb{X})$ of dimension n.

Table 1 shows some popular covariance functions used in the machine learning community [26], ordered by decreasing degree of smoothness, where θ is the correlation length parameter. These covariance functions are used in the numerical experiments of this paper.

3. Framework of the correspondence

Consider X as a nonempty set in $\mathbb{R}^d$, where $d \geq 1$, and let E be a Banach space of functions from X to $\mathbb{R}$. The smoothing problem is to estimate a function u defined on X that satisfies a priori information and a set of n noisy observations y_i at points $\boldsymbol{x}_i \in X$. In the Bayesian framework, Kimeldorf and Wahba in [14], chose the prior information to be a zero-mean GP $(U(\boldsymbol{x}))_{\boldsymbol{x}\in X}$, defined by its covariance function [26]

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}\left(U(\boldsymbol{x}), U(\boldsymbol{x}')\right) = \mathrm{E}\left[U(\boldsymbol{x})U(\boldsymbol{x}')\right], \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in X. \tag{3}$$

The Bayesian estimator $\tilde{u}$ of u is the mean of the posterior distribution of the GP $(U(\boldsymbol{x}))_{\boldsymbol{x}\in X}$, conditionally on the given data

$$\tilde{u}(\boldsymbol{x}) = \mathrm{E}\left[U(\boldsymbol{x}) \mid U(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\right]. \tag{4}$$

According to Section 2, the Bayesian estimation (4) has the following explicit expression [14, 26],

$$\tilde{u}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})^\top \left(\mathbb{K} + \sigma^2_{\mathrm{noise}}\mathbf{I}_n\right)^{-1} \boldsymbol{y}, \quad \boldsymbol{x} \in X,$$

where $\boldsymbol{k}(\boldsymbol{x}) = \left[k\left(\boldsymbol{x}, \boldsymbol{x}_1\right), \ldots, k\left(\boldsymbol{x}, \boldsymbol{x}_n\right)\right]^\top$ is the vector of covariance between $U(\boldsymbol{x})$ and $U(\mathbb{X})$, and $\mathbb{K} = k\left(\mathbb{X}, \mathbb{X}\right)$ is the covariance matrix of $U(\mathbb{X})$ of dimension $n \times n$.

For the unconstrained case (i.e., without any additional constraints on u), Kimeldorf and Wahba, in the late sixties, highlighted the correspondence between the Bayesian estimator $\tilde{u}$, defined by Equation (4), and the following

smoothing optimization problem [14]:

$$\min_{u\in H} \|u\|_H^2 + \frac{1}{\sigma_{\text{noise}}^2} \sum_{i=1}^{n} \left(u(\boldsymbol{x}_i) - y_i\right)^2, \tag{5}$$

where H is the RKHS associated with the GP $(U(\boldsymbol{x}))_{\boldsymbol{x}\in X}$, and σ_{noise} is the constant noise standard deviation.

Lemma 3.1 (Kimeldorf-Wahba correspondence [14])**.** *If U is a GP with the corresponding RKHS H, then*

$$\tilde{u} = \arg\min_{u\in H} \|u\|_H^2 + \frac{1}{\sigma_{noise}^2} \sum_{i=1}^{n} \left(u(\boldsymbol{x}_i) - y_i\right)^2,$$

where $\tilde{u}(\boldsymbol{x})$ is the Bayesian estimator (4), i.e., the mean of the posterior distribution $\{U(\boldsymbol{x}) \mid U(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$.

The notion of RKHS makes it possible to build a bridge between the deterministic world of optimization, Equation (5), and the probabilistic world of estimation, Equation (4). In 1950, Aronszajn published the theory of reproducing kernels [1], and in 1959, Parzen published 'Statistical inference on time series by Hilbert space methods' [25]. Later, Schwartz extended the formalism to topological spaces [29]. In the 1970s, Kimeldorf and Wahba popularized the use of reproducing kernel Hilbert spaces (RKHS) as a tool for providing efficient computations in Bayesian estimations [14, 4]. The GP $(U(\boldsymbol{x}))_{\boldsymbol{x}\in X}$ is represented by the Hilbert space spanned by the kernel k, which is defined by its covariance function (3). In the unconstrained case, the mean a posteriori (mAP) estimate, denoted as $\tilde{u}$, corresponds to the maximum a posteriori (MAP) estimate. This result is confirmed by the numerical example presented in Section 4.3.

We now consider the case where the underlying function u is known to satisfy additional information, such as monotonicity, boundedness, or convexity constraints [19, 21]. This can be expressed as $u \in C$, where C represents the set of functions satisfying the constraints. We can consider the mean of the posterior distribution,

$$\mathrm{E}\left[U(\boldsymbol{x}) \mid U \in C, U(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\right] \tag{6}$$

as in the unconstrained case. However, this conditional expectation (6) does not admit an explicit formula in general. The difficulty of making inference for the distribution of the GP U conditioned on the given data $\{U(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ and the constraints $\{U \in C\}$ lies in the fact that this conditional distribution is a truncated multivariate normal distribution in an infinite-dimensional linear space. As such, it cannot be expressed by a probability density function (pdf) associated with a Lebesgue equivalent measure.

To overcome this difficulty, we use the idea from approximation theory of approaching the original GP U with a finite-dimensional GP U_N that tends

to U in a certain sense as N tends to infinity (see [21] for more details). It makes sense to consider the density of the posterior distribution of the finite-dimensional GP

$$\{U_N(\boldsymbol{x}) \mid U_N \in C, U_N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\} \tag{7}$$

with respect to the N-dimensional Lebesgue measure. We define its MAP estimator as $\widehat{u}_N$, whose limit $\widehat{u}$ can be considered as an estimator of u. In Section 5, we prove that the MAP estimator $\widehat{u}_N$ of the finite-dimensional posterior distribution (7) is also the solution of the following discretized constrained optimization problem

$$\min_{u_N \in H_N \cap C} \|u_N\|^2_{H_N} + \frac{1}{\sigma^2_{\text{noise}}} \sum_{i=1}^{n} \left(u_N(\boldsymbol{x}_i) - y_i\right)^2, \tag{8}$$

where H_N is the RKHS associated with the GP approximation U_N and C is the set of functions verifying the constraints. Moreover, we prove that $\widehat{u}_N$ is convergent to $\widehat{u}$ solution of the following optimization problem

$$\min_{u \in H \cap C} \|u\|^2_{H} + \frac{1}{\sigma^2_{\text{noise}}} \sum_{i=1}^{n} \left(u(\boldsymbol{x}_i) - y_i\right)^2.$$

This result can be seen as a generalization of the Kimeldorf-Wahba correspondence (refer to Lemma 3.1) for constrained cases involving Bayesian estimation and optimal smoothing.

In what follows, for simplicity, we consider X to be the unit interval $[0, 1]$ in $\mathbb{R}$, and we take E to be the linear space of real-valued continuous functions on $[0, 1]$ equipped with the supremum norm, i.e., $E = \mathcal{C}([0, 1], \mathbb{R})$. The constraints are represented by a closed convex set C in E.

4. Constrained Gaussian processes

4.1. Constrained Gaussian process approximation

In this section, the parent GP U is approximated by a finite-dimensional GP U_N. To define U_N, a partition Δ_N of the interval $X = [0, 1]$ is required:

$$\Delta_N \ : \quad 0 = t_1 < \ldots < t_N = 1, \tag{9}$$

such that $\delta_N = \max\{|t_{j+1} - t_j|,\ j = 1, \ldots, N-1\}$ tends to zero as N tends to infinity. We assume

$$\Delta_N \subset \Delta_{N+1}. \tag{$H1$}$$

This assumption is not essential. It facilitates the proof of the convergence. By using the partition (9), we approximate the GP U with the following finite-dimensional GP [21, 19]:

$$U_N(x) := \sum_{j=1}^{N} U(t_j)\varphi_j(x) = \sum_{j=1}^{N} \xi_j \varphi_j(x), \quad x \in X, \tag{10}$$

where $[\varphi_1, \ldots, \varphi_N]^\top$ are the so-called hat functions, which are the basis of the piecewise linear function associated with the partition, and $\xi_j = U(t_j)$ for any $j \in 1, \ldots, N$. As far as we know, the basis sequence $\{\varphi_j\}$ originally proposed in [21] possesses several attractive properties, which are not necessarily shared by other bases, such as Bernstein polynomials [11], regression splines [6, 23], and restricted splines [30]. The hat functions are defined as follows:

$$\varphi_j(x) = \varphi\left(\frac{x - t_j}{\delta_N}\right), \quad x \in X, \tag{11}$$

for any $j = 1, \ldots, N$, where $\varphi(x) = (1 - |x|)\mathbb{1}_{[-1,1]}(x)$ is the *hat function* on $[-1, 1]$. Under Equation (10), the coefficients can be interpreted as the values of the GP $(U(x))_{x \in X}$ at the discretization points $\{t_j\}$. As U is a zero-mean GP, the vector $\boldsymbol{\xi} := [U(t_1), \ldots, U(t_N)]^\top$ is also zero-mean and Gaussian with a covariance matrix Γ_N:

$$\Gamma_N = \mathrm{Cov}(\boldsymbol{\xi}) = (\mathrm{Cov}(U(t_j), U(t_\ell)))_{1 \leq j, \ell \leq N} = (k(t_j, t_\ell))_{1 \leq j, \ell \leq N},$$

where we recall that k is the covariance function of the original GP U. These basis functions (11) have an interesting property: various shape constraints are *equivalent* to linear restrictions on the coefficient vector $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^\top$. For instance, let C be the set of functions that satisfy inequality constraints, such as boundedness, monotonicity, or convexity

$$C = \begin{cases} C_b := \left\{f \in \mathcal{C}^0(X, \mathbb{R}) \text{ s.t. } l_b \leq f(x) \leq u_b, \ \forall x \in X\right\} \\ C_m := \left\{f \in \mathcal{C}^0(X, \mathbb{R}) \text{ s.t. } f(x) \leq f(y), \ \forall x \leq y \in X\right\} \\ C_c := \left\{f \in \mathcal{C}^0(X, \mathbb{R}) \text{ s.t. } \frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(y)}{z - y}, \ \forall x \leq y \leq z \in X\right\} \end{cases} \tag{12}$$

where l_b and u_b are the lower and upper bounds respectively ($l_b < u_b \in \mathbb{R}$), and where $C^0(X, \mathbb{R})$ is the set of continuous functions from X to $\mathbb{R}$. Then,

$$U_N \in C \quad \Leftrightarrow \quad \boldsymbol{\xi} \in \mathcal{E} \tag{13}$$

where $\mathcal{E}$ is a set of linear constraints on the coefficients $\boldsymbol{\xi}$ in $\mathbb{R}^N$. According to (12), we get

$$\mathcal{E} = \begin{cases} \mathcal{E}_b := \left\{z \in \mathbb{R}^N : \ l_b \leq z_j \leq u_b, \ \forall j = 1, \ldots, N\right\} \\ \mathcal{E}_m := \left\{z \in \mathbb{R}^N : \ z_{j-1} \leq z_j, \ \forall j = 2, \ldots, N\right\} \\ \mathcal{E}_c := \left\{z \in \mathbb{R}^N : \ \frac{z_{j-1} - z_{j-2}}{t_{j-1} - t_{j-2}} \leq \frac{z_j - z_{j-1}}{t_j - t_{j-1}}, \ \forall j = 3, \ldots, N\right\} \end{cases} \tag{14}$$

which corresponds to boundedness, monotonicity, and convexity constraints respectively. Let us mention that the above linear inequality constraints on the coefficients (14) can be written in a matrix form as follows:

$$\left\{\boldsymbol{\xi} \in \mathbb{R}^N \text{ s.t. } \boldsymbol{l} \leq \Lambda\boldsymbol{\xi} \leq \boldsymbol{u}\right\},$$

where $\Lambda \in \mathbb{R}^{m \times N}$ is the matrix of constraints, and $\boldsymbol{l}$ and $\boldsymbol{u}$ are respectively the lower and upper bounds vectors. For the boundedness constraints $\boldsymbol{\xi} \in \mathcal{E}_b$, the

matrix of constraints Λ is the $N \times N$ identity matrix $\mathbf{I}_N$, and the lower and upper bounds vectors are $\boldsymbol{l} = [l_b, \ldots, l_b]^\top \in \mathbb{R}^N$ and $\boldsymbol{u} = [u_b, \ldots, u_b]^\top \in \mathbb{R}^N$ respectively. In that case, sampling the GP approximation U_N under linear inequality constraints $\{U_N \in \mathcal{C}\}$ is *equivalent* to the simulation of the Gaussian vector $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_N, \Gamma_N)$ such that $\boldsymbol{l} \leq \Lambda\boldsymbol{\xi} \leq \boldsymbol{u}$ [20, 24].

4.2. Constrained Gaussian process approximation with noisy observations

In this section, the GP approximation (10) with both noisy observations and linear inequality constraints is considered

$$U_N(x) = \sum_{j=1}^{N} \xi_j \varphi_j(x), x \in X \quad \text{s.t.} \quad \begin{cases} U_N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y} & \text{(noisy observations)}, \\ U_N \in C & \text{(inequality constraints)}, \end{cases}$$

where $\boldsymbol{\epsilon}$ is an additive and independent zero-mean Gaussian vector with noise covariance matrix $\sigma^2_{\text{noise}}\mathbf{I}_n$. The noisy observations can be written in a matrix form as follows:

$$\mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y},$$

where $\mathbf{A}$ is the $n \times N$ matrix defined as $\mathbf{A}_{i,j} := \varphi_j(x_i)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, N$. Based on the equivalence shown in (13), we can obtain the conditional distribution of U_N under both noisy observations $\{U_N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ and constraints $\{U_N \in C\}$ from the conditional distribution of the Gaussian coefficients $\boldsymbol{\xi}$ given $\{\mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ and $\{\boldsymbol{\xi} \in \mathcal{E}\}$

$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_N, \Gamma_N) \quad \text{s.t.} \quad \begin{cases} \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}, \\ \boldsymbol{\xi} \in \mathcal{E}. \end{cases} \tag{15}$$

The sampling procedure is summarized as follows: under only noisy observations, the conditional distribution of $\boldsymbol{\xi}$ is a multivariate normal (MVN) [22, 26]

$$\{\boldsymbol{\xi} \mid \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{where,}$$

$$\begin{cases} \boldsymbol{\mu} &= (\mathbf{A}\Gamma_N)^\top (\mathbf{A}\Gamma_N\mathbf{A}^\top + \sigma^2_{\text{noise}}\mathbf{I}_n)^{-1}\boldsymbol{y}, \\ \boldsymbol{\Sigma} &= \Gamma_N - (\mathbf{A}\Gamma_N)^\top (\mathbf{A}\Gamma_N\mathbf{A}^\top + \sigma^2_{\text{noise}}\mathbf{I}_n)^{-1}\mathbf{A}\Gamma_N. \end{cases}$$

Thus, the posterior distribution (15) is the following truncated MVN

$$\{\boldsymbol{\xi} \mid \boldsymbol{\xi} \in \mathcal{E}, \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\} \sim \mathcal{N}_{\mathcal{T}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{l}, \boldsymbol{u}),$$

where $\mathcal{N}_{\mathcal{T}}(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{a}, \boldsymbol{b})$ is the MVN distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{M}$ truncated between the lower bound $\boldsymbol{a}$ and upper bound $\boldsymbol{b}$. In this paper, the efficient Hamiltonian Monte Carlo (HMC) sampler [24] implemented in the R package *tmg* is used. Now, we can define the MAP estimate of the GP approximation (10). Before defining the MAP estimate, we compute the

posterior mode $\boldsymbol{\mu}^*$ of the posterior distribution (15), which corresponds to the maximum of the posterior pdf,

$$\boldsymbol{\mu}^* := \arg\min_{\boldsymbol{\xi}} \left\{ \boldsymbol{\xi}^\top \Gamma_N \boldsymbol{\xi} \;\middle|\; \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{l} \leq \Lambda\boldsymbol{\xi} \leq \boldsymbol{u} \right\}. \tag{16}$$

This is a quadratic optimization problem under linear inequality constraints [5, 13]. In this paper, the function solve.QP of the R package *quadprog* is used to compute $\boldsymbol{\mu}^*$ in (16).

Definition 4.1 (MAP estimate). The Maximum a posterior (MAP) estimate of U_N conditionally on the inequality constraints and noisy observations is defined as

$$\widehat{u}_N(x) := \sum_{j=1}^{N} \boldsymbol{\mu}_j^* \varphi_j(x) = \varphi(x)^\top \boldsymbol{\mu}^*, \quad x \in X, \tag{17}$$

where $\boldsymbol{\mu}^* \in \mathbb{R}^N$ is the posterior mode defined in (16) and $\varphi(x) = [\varphi_1(x), \ldots, \varphi_N(x)]^\top$.

Note that the MAP estimate is independent of the sampling procedure and only depends on a quadratic optimization problem with linear inequality constraints (16).

Definition 4.2 (mAP estimate). The mean a posteriori (mAP) estimate of U_N conditionally on the inequality constraints and noisy observations is defined as

$$\tilde{u}_N(x) := \mathrm{E}\left[U_N(x) \mid U_N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}, U_N \in C\right] = \varphi(x)^\top \boldsymbol{\mu}_c, \tag{18}$$

where $\boldsymbol{\mu}_c := \mathrm{E}\left[\boldsymbol{\xi} \mid \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{l} \leq \Lambda\boldsymbol{\xi} \leq \boldsymbol{u}\right]$ is the posterior mean which is computed from simulations.

In the next section, the behavior of these two estimates (MAP and mAP) is investigated. Additionally, a comparison in terms of mean squared prediction error (MSPE) is given. The MAP estimate produces better results in terms of MSPE compared to the mAP estimate.

4.3. Motivating example

In this section, we examine the behavior of the MAP and mAP estimates using a bounded function defined on $X = [0, 1]$, as follows

$$f_1(x) = \begin{cases} \cos\left(\pi(2x + \frac{1}{3})\right) & \text{if } x \in [0, 2/3]; \\ 0.5 & \text{otherwise.} \end{cases}$$

This is a challenging situation because the true function, which is bounded between -1 and 0.5, is flat in the region $[2/3, 1]$ and attains the upper bound constraint.

Figure 1 shows the GP approach (10) with boundedness constraints. We used the efficient HMC sampler [24] to sample from the posterior distribution of the basis coefficients $\{\xi_j\}$ in (15). We fix $N = \lfloor n/8 \rfloor$, where n is the number of

training sample. We use the Matérn covariance function with regularity parameter $\nu = 5/2$ (see Table 1), and correlation length parameter θ fixed at 0.4. The black stars represent the 100 training data points generated randomly from (1) with true function f_1 and a true noise variance $\sigma^2_{\text{noise}} = 0.4^2$. The black solid curve represents the true bounded function f_1, while the red dashed curve corresponds to the MAP estimate $\widehat{u}_N$ (17), and the blue dashed-dotted curve corresponds to the mAP estimate $\tilde{u}_N$ (18). The black dashed horizontal lines represent the lower and upper bound constraints. The gray shaded area represents the 95% pointwise confidence interval. Both the mAP and MAP estimates satisfy the boundedness constraints over the entire domain. In the left panel, the MAP estimate fits the model better, especially where the true function is close to the lower or upper bounds, whereas the mAP estimate fails to capture the flat region. To rigorously compare the two estimates, we put a prior on the correlation length parameter $\theta \sim \mathcal{U}(0.3, 1)$, and generate a dataset of size $n = 500$ from (1) using the true function f_1 and a true standard deviation $\sigma_{\text{noise}} = 0.4$. They are randomly split into a training set of size 300 and a testing set of size 200. We conduct numerical studies on 1,000 replicates, obtaining an average MSPE of 7.41×10^{-3} (standard deviation 5.19×10^{-3}) for the MAP estimate and 9.19×10^{-3} (standard deviation 5.03×10^{-3}) for the mAP estimate.

In the right panel, we relaxed the boundedness constraints from $[-1, 0.5]$ to $[-1.5, 1]$, a range where the constraints do not impact the model. As expected, the MAP and mAP estimates coincide, confirming that the mAP estimate for the unconstrained case is also the MAP.

5. Main results: asymptotic analysis

In this section, we investigate the asymptotic behavior of model (10) and establish the convergence of the MAP estimator $\widehat{u}_N$ to the constrained optimal smoothing solution of problem (P) through rigorous analysis. The covariance function of the GP approximation U_N can be computed as follows:

$$\forall x, x' \in X, \quad k_N(x, x') = \sum_{j,\ell=1}^{N} k(t_j, t_\ell)\varphi_j(x)\varphi_\ell(x') = \varphi(x)^\top \Gamma_N \varphi(x'),$$

where $\varphi(x) = [\varphi_1(x), \ldots, \varphi_N(x)]^\top$, for any $x \in X$. The GP approximation U_N induces a RKHS H_N which is the classical subspace of piecewise linear continuous functions associated to the partition Δ_N:

$$H_N := \text{Span}\{\varphi_j,\ j = 1, \ldots, N\}.$$

It is equipped with the following scalar product: for all $u_N, v_N \in H_N$

$$\langle u_N, v_N \rangle_{H_N} := c_{u_N}^\top \Gamma_N^{-1} c_{v_N},$$

where $c_f := [f(t_1), \ldots, f(t_N)]^\top$. This scalar product induces the following norm

$$\|u_N\|^2_{H_N} = c_{u_N}^\top \Gamma_N^{-1} c_{u_N}.$$

In the following proposition, we prove that H_N is a RKHS with kernel k_N associated with U_N, and is a Hilbertian subspace of E.

Lemma 5.1 (Characterization of H_N)**.** *As in Theorem 1 of [3] and Lemma 2 of [2], H_N is a RKHS with kernel k_N associated with U_N, and is a Hilbertian subspace of E, that is, for any $h_N \in H_N$,*

$$\|h_N\|_E \leq d\|h_N\|_{H_N}, \tag{19}$$

where d is a constant independent of N.

Proof. The proof is given in [2], Lemma 2, p. 1592. □

The posterior pdf of $\{U_N \mid U_N \in C, U_N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ is of the form

$$L^N_{pos}(u_N) = \nabla_N^{-1} \mathbb{1}_{u_N \in H_N \cap C} \exp\left(-\frac{1}{2} J_N(u_N)\right),$$

where ∇_N is a normalizing constant and

$$J_N(u_N) := \|u_N\|^2_{H_N} + \frac{1}{\sigma^2_{\text{noise}}} \sum_{i=1}^{n} (u_N(x_i) - y_i)^2.$$

Then, the MAP estimator $\widehat{u}_N$ (17) as the mode of the posterior distribution $\{U_N \mid U_N \in C, U_N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ is well defined and is the solution of the discretized optimization problem

$$\min_{u_N \in H_N \cap C} \|u_N\|^2_{H_N} + \frac{1}{\sigma^2_{\text{noise}}} \sum_{i=1}^{n} (u_N(x_i) - y_i)^2,$$

which can be rewritten as

$$\min_{u_N \in H_N \cap C} J_N(u_N). \tag{P_N}$$

The problem (P_N) can be seen as the discretization of the *constrained optimal smoothing*

$$\min_{u \in H \cap C} J(u), \quad \text{where} \tag{P}$$

$$J(u) := \|u\|^2_H + \frac{1}{\sigma^2_{\text{noise}}} \sum_{i=1}^{n} (u(x_i) - y_i)^2.$$

Before showing the existence and the uniqueness of problems (P_N) and (P), let us give the following two assumptions.

1. Assume that
$$\pi_N(C) \subset C, \tag{$H2$}$$
where π_N is the classical piecewise linear interpolation projection defined from E onto H_N by
$$\forall f \in E, \quad \pi_N(f) = \sum_{j=1}^{N} f(t_j)\varphi_j.$$

Hypothesis (H2) is clearly fulfilled in the cases treated in this paper (boundedness and monotonicity constraints).

2. The set C is a closed convex set of E such that

$$H \cap C \neq \emptyset. \tag{H3}$$

This assumption is important, in particular, in the proof of the existence of problems (P) and (P_N).

Proposition 1 (Uniqueness of problem (P_N)). Under hypotheses (H2) and (H3), the problem (P_N) has a unique solution denoted $\widehat{u}_N$.

Proof. J_N is strongly convex, Fréchet differentiable and

$$\lim_{\|v_N\|_{H_N} \to +\infty} J_N(v_N) = +\infty.$$

Let $g \in H \cap C$, then, thanks to hypothesis (H2), $\pi_N(g) \in H_N \cap C$. So that $H_N \cap C$ is a nonempty closed convex set of H_N. □

Proposition 2 (Uniqueness of problem (P)). Under (H3), the problem (P) has a unique solution, denoted by $\widehat{u}$.

Proof. It is easy to see that J is Fréchet differentiable and

$$\lim_{\|v\|_H \to +\infty} J(v) = +\infty.$$

Moreover, J is strongly convex: for all $u, v \in H$ and $t \in [0,1]$,

$$J(tu + (1-t)v) \leq tJ(u) + (1-t)J(v) - t(1-t)\|u - v\|_H^2.$$

As $H \cap C$ is a nonempty closed convex set, we have the result. □

Now, we are ready to give the main result of this paper which is represented by the following important theorem.

Theorem 5.2 (Convergence to the constrained optimal smoothing)**.** *Under hypotheses (H1) and (H2), we have the following result: If $\widehat{u}$ is the solution of problem (P), then*

$$\widehat{u}_N \underset{N \to +\infty}{\longrightarrow} \widehat{u} \quad \text{in } E,$$

where we recall that E is the usual Banach space of continuous functions on $X = [0,1]$.

Before showing the proof of Theorem 5.2, we give some intermediate results represented by series of propositions with their proofs. They are needed to show that, if $\widehat{u}_N$ is the solution of (P_N) and $\widehat{u}$ is the solution of (P), then in E,

$$\lim_{N \to +\infty} \widehat{u}_N = \widehat{u}.$$

This means that the MAP estimate using the proposed approach, $\widehat{u}_N$, converges uniformly to the *constrained optimal smoothing* solution, $\widehat{u}$, of problem (P). The proof is highly technical and involves approximation theory and functional analysis, including weak convergence and compactness.

Proposition 3 (Stability of π_N). For any $h \in H$, we have

$$\|\pi_N(h)\|_{H_N} \leq \|h\|_H. \tag{20}$$

Moreover, H is characterized by

$$H = \left\{ h \in E \ : \ \sup_N \|\pi_N(h)\|_{H_N} < +\infty \right\} \tag{21}$$

and, for all $h \in H$, by

$$\|h\|_H^2 = \lim_{N \to +\infty} \|\pi_N(h)\|_{H_N}^2. \tag{22}$$

Proof. See [2], Theorem 3.1, p. 1587. □

Let us define the linear operator $\rho_N : H_N \to H$ defined by

$$\forall v_N \in H_N, \quad \rho_N(v_N) := \sum_{j=1}^{N} \theta_j k(\cdot, t_j), \tag{23}$$

where $\Theta = [\theta_1, \ldots, \theta_N]^\top$, solves $\Gamma_N \Theta = c_{v_N}$.

Proposition 4 (Isometric property of ρ_N and convergence property). For all $v_N \in H_N$,

$$\|\rho_N(v_N)\|_H^2 = c_{v_N}^\top \Gamma_N^{-1} c_{v_N}. \tag{24}$$

The operator ρ_N is an isometry from H_N into H, i.e.

$$\forall v_N \in H_N, \qquad \|\rho_N(v_N)\|_H^2 = \|v_N\|_{H_N}^2. \tag{25}$$

$$\forall h \in H, \qquad \|\rho_N(\pi_N(h)) - h\|_H \underset{N \to +\infty}{\longrightarrow} 0. \tag{26}$$

Proof. The proof is detailed in [3], Propositions 3 and 5. □

Proposition 5. We have

$$\sup_{x \in X} \|\rho_N(k_N(\cdot, x)) - k(\cdot, x)\|_H \underset{N \to +\infty}{\longrightarrow} 0.$$

Proof. The proof is given in [3], Lemma 5. □

We will prove two intermediate results leading to the convergence result.

Proposition 6. Under hypotheses (*H*1), (*H*2) and (*H*3), we get

$$\lim_{N \to +\infty} \quad J_N(\pi_N(\widehat{u})) = J(\widehat{u});$$

$$\lim_{N \to +\infty} \quad J_N(\widehat{u}_N) = J(\widehat{u}).$$

The proof of Proposition 6 is deferred to the Appendix.

Proposition 7. We have the following result

$$\lim_{N \to +\infty} \|\pi_N(\widehat{u}) - \widehat{u}_N\|_{H_N} = 0. \tag{27}$$

Proof. As J_N is strongly convex and differentiable, then

$$J_N(\pi_N(\widehat{u})) - J_N(\widehat{u}_N) \geq \langle J'_N(\widehat{u}_N), \pi_N(\widehat{u}) - \widehat{u}_N \rangle_{H_N} + \|\pi_N(\widehat{u}) - \widehat{u}_N\|^2_{H_N},$$

where J'_N denotes the derivative of J_N. As $\pi_N(\widehat{u}) \in H_N \cap C$ and $\widehat{u}_N$ solves (P_N),

$$\langle J'_N(\widehat{u}_N), \pi_N(\widehat{u}) - \widehat{u}_N \rangle_{H_N} \geq 0,$$

so that, $\|\pi_N(\widehat{u}) - \widehat{u}_N\|^2_{H_N} \leq J_N(\pi_N(\widehat{u})) - J_N(\widehat{u}_N)$. The result in (27) comes from the application of Proposition 6. □

Proof of Theorem 5.2. We have

$$\|\widehat{u}_N - \widehat{u}\|_E \leq \|\widehat{u}_N - \pi_N(\widehat{u})\|_E + \|\pi_N(\widehat{u}) - \widehat{u}\|_E.$$

We know from approximation theory in the Banach space E that

$$\|\pi_N(\widehat{u}) - \widehat{u}\|_E \underset{N \to +\infty}{\longrightarrow} 0. \tag{28}$$

As H_N is a Hilbertian subspace of E, see Equation (19),

$$\|\pi_N(\widehat{u}) - \widehat{u}_N\|_E \leq c\|\pi_N(\widehat{u}) - \widehat{u}_N\|_{H_N}.$$

Proposition 7 gives the result. □

6. Performance behavior of the MAP

In this section the correspondence established in the previous section between the MAP estimate and the optimal *constrained* smoothing function solution of problem (P) is illustrated. Additionally, the performance of the proposed approach (10) is investigated and the advantage of the MAP estimate $\widehat{u}_N$ (17) over the mAP estimate $\tilde{u}_N$ (18) in terms of prediction accuracy is highlighted. We consider the case where the function f satisfies monotonicity (nondecreasing) constraints. In this case, the associated convex set is given by :

$$C = \left\{ f \in \mathcal{C}^0\left([0,1]\right) \ : \ f(x) \leq f(x'), \forall x \leq x' \right\}.$$

The approximately flat function used in [33] is considered:

$$f_2(x) = \sqrt{2} \sum_{\ell=1}^{100} \ell^{-0.7} \sin(\ell) \cos(\pi(\ell - 0.5)(1 - x)), \quad \forall x \in [0,1]. \tag{29}$$

This is a challenging situation because the real function f_2 is slightly flat on $[0.7, 1]$ and it decreases in certain regions, which allows us to evaluate the performance of the proposed model under slight model misspecification. The aim is to investigate the behavior of both the MAP and mAP estimates in terms of prediction accuracy.

In Figure 2, the GP approach developed in this paper (10) has been used to assess the prediction accuracy of the function f_2 defined in (29). We fix $N = \lfloor n/8 \rfloor$, where n is the number of training samples. We use the Matérn covariance function with regularity parameter $\nu = 3/2$ (Table 1), and correlation length parameter θ fixed at 0.4. The efficient HMC sampler from the R package *tmg* was used to sample from the posterior distribution of the basis coefficients $\{\xi_j\}$ (15). We generate 500 pairs of responses from (1) using the true function f_2 and a true noise variance $\sigma^2_{\text{noise}} = 0.4^2$. The dataset is divided randomly into training samples of size 300 and testing samples of size 200. The black stars in Figure 2 represent the 300 training samples. The black curve corresponds to the true function, the red dashed curve represents the MAP estimate, and the blue dashed-dotted curve corresponds to the mAP estimate. The mAP estimate $\tilde{u}_N$ (18) was computed using $5,000$ samples of the posterior distribution obtained with the efficient HMC sampler. We observe that the posterior distribution and the mAP estimate fail to capture the flat region, contrary to the MAP estimate. This is due to the phenomenon of *mass-shifting*, which is well described in [33].

To rigorously compare the MAP and mAP estimates in terms of prediction accuracy, we propose to put a prior on the correlation length parameter $\theta \sim \mathcal{U}(0.3, 1)$. We generate a dataset of size 500 from (1) using the true function f_2 and a true noise variance $\sigma^2_{\text{noise}} = 0.4^2$ and randomly split it into a training set of size 300 and a testing set of size 200. When using the MAP estimate, we get an average MSPE over 1,000 replicates equal to 4.01×10^{-3} (standard deviation 1.45×10^{-3}), and when using the mAP estimate, we get an average MSPE over 1,000 replicates equal to 7.04×10^{-3} (standard deviation 1.68×10^{-3}).

The monotonicity constraints in two dimensions is considered. Without loss of generality, we suppose that X^2 is the unit square, i.e., $X^2 = [0, 1]^2$. The real function f is supposed monotone (nondecreasing) with respect to the two input variables. For any $x = (x_1, x_2)$ and $x' = (x'_1, x'_2)$ in X^2, we have

$$x_1 \leq x'_1 \quad \text{and} \quad x_2 \leq x'_2 \quad \Rightarrow \quad f(x_1, x_2) \leq f(x'_1, x'_2).$$

The two-dimensional GP approximation is defined as follows:

$$U_N(x_1, x_2) := \sum_{j=1}^{N_1} \sum_{\ell=1}^{N_2} U(t_j, t_\ell) \varphi^1_j(x_1) \varphi^2_\ell(x_2), \quad \forall (x_1, x_2) \in X^2, \tag{30}$$

where φ^1_j and φ^2_ℓ refer to the hat functions defined in the one-dimensional case (11), and N_1 and N_2 represent the number of discretization points (i.e., knots) for the corresponding input sets.

In Figure 3, the two-dimensional squared exponential (SE) covariance function has been used:

$$k(x, x') = \exp\left(-\frac{(x_1 - x'_1)^2}{2\theta_1^2} - \frac{(x_2 - x'_2)^2}{2\theta_2^2} \right),$$

where $x = (x_1, x_2)$ and $x' = (x'_1, x'_2)$ are in $X^2 = [0, 1]^2$, and the correlation length parameters (θ_1, θ_2) are fixed at $(0.5, 0.8)$. We set $N_1 = N_2 = 7$, resulting

in 49 basis functions and knots. The five hundred training data points (represented by black stars) were generated using the Hypercube Latin method from the R package *lhs*, as well as from the regression problem given in (1), where the true nondecreasing function is $f(x_1, x_2) = 3/\left(1 + \exp(-10x_1 + 0.2)\right) + x_2 + 2$ and the true noise variance is $\sigma^2_{\text{noise}} = 0.4^2$. In the left panel, we illustrate the true function f together with the training data. In the middle panel, we show the mAP estimate $\tilde{u}_N$, which is computed using the HMC sampler. The right panel presents the MAP estimate $\widehat{u}_N$ alongside the training data.

To rigorously compare the accuracy of the mAP and MAP estimates in terms of prediction, we propose to put a prior on the correlation length parameters θ_1 and θ_2, with $\theta_1 \sim \mathcal{U}(0.1, 1)$ and $\theta_2 \sim \mathcal{U}(0.1, 1)$, as well as on the noise standard deviation σ_{noise}, with $\sigma_{\text{noise}} \sim \mathcal{U}(0.5, 1)$. We conduct a numerical experiment with one hundred replicates, using a dataset of size $n = 500$ generated from (1) with the true function f and a noise standard deviation σ_{noise}. The dataset is randomly split into 80% training samples and 20% testing samples. In this case, the average MSPE is equal to 2.68×10^{-2} with standard deviation of order 1.08×10^{-2} when using the MAP estimate and equal to 3.07×10^{-2} with standard deviation of order 1.12×10^{-2} when using the mAP estimate.

7. Real application (age and income dataset)

In this section, the GP approach (10) has been applied to a real-life dataset consisting of age (in years) and the logarithm of income (log.income) for 205 Canadian workers from a 1971 Canadian Census Public Use Tape. The aim is to estimate the logarithm of income as function of age. This real-life data will be used to demonstrate the performance of the MAP compared to the mAP estimate in terms of prediction accuracy. Data suggests that the underlying function is monotone nondecreasing with a flat region when `age` is greater than 26. The Matérn covariance function with regularity parameter $\nu = 5/2$ is used as recommended in [26]. We compare the performance of the MAP and mAP estimates in terms of prediction accuracy through the MSPE criterion.

Figure 4 presents performance illustrations of the proposed approach (10) for the real-world application. We set $N = \lfloor n/8 \rfloor$ to avoid overfitting. This choice is justified later in this section. We use $5,000$ samples obtained by the efficient HMC technique to get the mAP estimate $\tilde{u}_N$ (18) as well as the 95% pointwise credible interval (gray shaded region). The red dashed curve represents the MAP estimate, while the blue dashed-dotted curve represents the mAP estimate. We observe that both the MAP and mAP estimates align the data points well. However, contrary to the MAP estimate, the posterior distribution as well as the mAP estimate fail to capture the flat region ($\texttt{age} \geq 26$). This is because of the *mass-shifting* phenomenon that arises from the truncated MVN sampling, as described in [33]. To compare the two estimates MAP and mAP rigorously in terms of prediction accuracy, we propose putting a prior on the correlation length parameter, $\theta \sim \mathcal{U}(10, 50)$, and on the noise standard deviation, $\sigma_{\text{noise}} \sim \mathcal{U}(0.5, 1)$. By randomly splitting the total dataset of size 205 into 80%

training and 20% testing datasets, we obtain an average MSPE over one thousand replicates of 33.84×10^{-2} when using the MAP estimate and 36.82×10^{-2} when using the mAP estimate.

To avoid overfitting, we analyze the value of the number of discretization points, N, as a function of the number of samples, since its value influences the prediction accuracy of the proposed approach (10). It is more reasonable to choose the number of discretization points, N, to be smaller than the number of samples. We consider the case where $N \in \{n, \lfloor n/2 \rfloor, \lfloor n/4 \rfloor, \lfloor n/8 \rfloor\}$ to conduct a thorough analysis on our real-world dataset application, where n represents the number of training samples.

Figure 5 shows the 5-fold CV MSPEs repeated fifty times as a function of the correlation length parameter, $\theta \in \{5, 10, 20, 30, 40\}$, with $N \in \{n, \lfloor n/2 \rfloor, \lfloor n/4 \rfloor, \lfloor n/8 \rfloor\}$. The noise standard deviation is fixed at $\sigma_{\text{noise}} = 0.5$ (resp. $\sigma_{\text{noise}} = 1$) in the left (resp. right) panel. First, we observe that in both cases, the MSPEs drop rapidly for small values of the correlation length parameter θ and then increase for large values of θ. Second, we observe that using a different number of discretization points N provides similar MSPE values. These numerical experiments guided us to choose, for example, $N = \lfloor n/8 \rfloor$, in this real-world application. Finally, we mention that with the smallest standard deviation σ_{noise}, the optimal MSPE is obtained at a smaller correlation length parameter, whereas with the largest standard deviation, the optimal MSPE is obtained at a larger correlation length parameter. In fact, the noise standard deviation σ_{noise} can be interpreted as a compromise between smoothness and fidelity to the data.

8. Conclusion and applications

The purpose of this paper was to address the problem of inferring a GP denoted U based on noisy observations and a set of constraints represented by a closed convex set C. To do so, we first approximated the GP U with an efficient finite-dimensional GP U_N. We then demonstrated that the Maximum a Posteriori (MAP) of the conditional distribution (posterior distribution) corresponds to a constrained discrete optimal smoothing problem, which is convergent to a constrained optimal smoothing problem in the Reproducing Kernel Hilbert Space (RKHS) associated with U as the dimension N tends to infinity. This result generalizes the correspondence established by Kimeldorf and Wahba [14]. We also showed that for constrained cases, the optimal smoothing corresponds to the MAP and not to the mean a posteriori (mAP) estimate. Both synthetic and real datasets were used to confirm the theoretical result.

Proof of Proposition 6

Proof of Proposition 6. Let us set

$$h^N := \rho_N(\widehat{u}_N) \in H.$$

Let us prove that there exists $\varepsilon_N \to 0$ and $\eta_N \to 0$ as $N \to +\infty$ and constants c_1 and c_3 such that

$$J(h^N) \leq J_N(\widehat{u}_N) + c_3\varepsilon_N \leq J_N(\pi_N(\widehat{u})) + c_3\varepsilon_N \leq J(\widehat{u}) + c_1\eta_N + c_3\varepsilon_N. \quad (31)$$

According to (19) and (20),

$$\begin{aligned} |\pi_N(\widehat{u})(x_i)| &\leq \|\pi_N(\widehat{u})\|_E \leq d\|\pi_N(\widehat{u})\|_{H_N} \leq d\|\widehat{u}\|_H \\ |\widehat{u}(x_i)| &\leq \|\widehat{u}\|_E \leq c\|\widehat{u}\|_H. \end{aligned}$$

Hence

$$\begin{aligned} J_N(\pi_N(\widehat{u})) &= \|\pi_N(\widehat{u})\|^2_{H_N} + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}(\pi_N(\widehat{u})(x_i) - y_i)^2 \\ &\leq \|\widehat{u}\|^2_H + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}(\pi_N(\widehat{u})(x_i) - y_i)^2 \\ &\leq J(\widehat{u}) + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left[(\pi_N(\widehat{u})(x_i) - y_i)^2 - (\widehat{u}(x_i) - y_i)^2\right] \\ &\leq J(\widehat{u}) + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left[\pi_N(\widehat{u})(x_i) - \widehat{u}(x_i)\right]\left[\widehat{u}(x_i) + \pi_N(\widehat{u})(x_i) - 2y_i\right] \\ J_N(\pi_N(\widehat{u})) &\leq J(\widehat{u}) + \underbrace{\frac{n\left[(c+d)\|\widehat{u}\|_E + 2\max|y_i|\right]}{\sigma^2_{\text{noise}}}}_{c_1}\|\pi_N(\widehat{u}) - \widehat{u}\|_E. \end{aligned}$$

Thus, setting $\eta_N = \|\pi_N(\widehat{u}) - \widehat{u}\|_E$, according to (28), we have $\eta_N \to 0$ as $N \to +\infty$.

$$J_N(\pi_N(\widehat{u})) \leq J(\widehat{u}) + c_1\eta_N,$$

where the constant c_1 is independent of N. By definition of $\widehat{u}_N$, $J_N(\widehat{u}_N) \leq J_N(\pi_N(\widehat{u}))$ so that

$$J_N(\widehat{u}_N) \leq J(\widehat{u}) + c_1\eta_N.$$

We have

$$|\rho_N(\widehat{u}_N)(x_i)| \leq \|\rho_N(\widehat{u}_N)\|_E \leq c\|\rho_N(\widehat{u}_N)\|_H = c\|\widehat{u}_N\|_{H_N} \leq cJ_N(\widehat{u}_N) \leq cJ(\widehat{u}) + cc_1\eta_N$$

and

$$|\widehat{u}_N(x_i)| \leq \|\widehat{u}_N\|_E \leq d\|\widehat{u}_N\|_{H_N} \leq dJ_N(\widehat{u}_N) \leq dJ(\widehat{u}) + dc_1\eta_N.$$

Let c_2 be such that, for all N,

$$(c+d)(J(\widehat{u}) + c_1\eta_N) \leq c_2.$$

We have, according to the isometric property of ρ_N,

$$\begin{aligned}
J(\rho_N(\widehat{u}_N)) &= \|\rho_N(\widehat{u}_N)\|_H^2 + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left(\rho_N(\widehat{u}_N)(x_i)-y_i\right)^2 \\
&= \|\widehat{u}_N\|_{H_N}^2 + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left(\rho_N(\widehat{u}_N)(x_i)-y_i\right)^2 \\
&= J_N(\widehat{u}_N) - \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left(\widehat{u}_N(x_i)-y_i\right)^2 + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left(\rho_N(\widehat{u}_N)(x_i)-y_i\right)^2 \\
&= J_N(\widehat{u}_N) + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left[\left(\rho_N(\widehat{u}_N)(x_i)-y_i\right)^2 - \left(\widehat{u}_N(x_i)-y_i\right)^2\right] \\
&= J_N(\widehat{u}_N) + \frac{1}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left[\rho_N(\widehat{u}_N)(x_i)+\widehat{u}_N(x_i)-2y_i\right]\left[\rho_N(\widehat{u}_N)(x_i)-\widehat{u}_N(x_i)\right] \\
&\leq J_N(\widehat{u}_N) + \frac{c_2+2\max|y_i|}{\sigma^2_{\text{noise}}}\sum_{i=1}^{n}\left[\rho_N(\widehat{u}_N)(x_i)-\widehat{u}_N(x_i)\right].
\end{aligned}$$

Thanks to the reproducing properties of H and H_N and to the isometric property of ρ_N:

$$\begin{aligned}
|\rho_N(\widehat{u}_N)(x_i)-\widehat{u}_N(x_i)| &= |\langle\rho_N(\widehat{u}_N),k(\cdot,x_i)\rangle_H - \langle\widehat{u}_N,k_N(\cdot,x_i)\rangle_{H_N}| \\
&= |\langle\rho_N(\widehat{u}_N),k(\cdot,x_i)\rangle_H - \langle\rho_N(\widehat{u}_N),\rho_N(k_N(\cdot,x_i))\rangle_H| \\
&= |\langle\rho_N(\widehat{u}_N),k(\cdot,x_i)-\rho_N(k_N(\cdot,x_i))\rangle_H| \\
&\leq \|\rho_N(\widehat{u}_N)\|_H \times \|k(\cdot,x_i)-\rho_N(k_N(\cdot,x_i))\|_H \\
&\leq (J(\widehat{u})+c_1\eta_N)\max_i\|k(\cdot,x_i)-\rho_N(k_N(\cdot,x_i))\|_H.
\end{aligned}$$

Let us set

$$\varepsilon_N = \max_i\|k(\cdot,x_i)-\rho_N(k_N(\cdot,x_i))\|_H.$$

According to Proposition 5, $\varepsilon_N \to 0$ as $N \to +\infty$ and

$$J(\rho_N(\widehat{u}_N)) \leq J_N(\widehat{u}_N) + c_3\varepsilon_N, \quad \text{with}$$

$$c_3 = \frac{nc_2+2n\max|y_i|}{\sigma^2_{\text{noise}}}.$$

Then, we get (31). As

$$\|h^N\|_H \leq J(h^N) \leq J(\widehat{u}) + c_1\eta_N + c_3\varepsilon_N,$$

then the sequence $(h^N)_{N\in\mathbb{N}}$ is bounded in H so that, by weak compactness in the Hilbert space, there exists a sub-sequence $(h^{N_k})_{k\in\mathbb{N}}$ and $h^* \in H$ such that

$$h^{N_k} \underset{k\to+\infty}{\rightharpoonup} h^* \in H, \qquad \text{(weak convergence)}.$$

As H is a RKHS with kernel K, for all $t_j \in \Delta_N, k(\cdot, t_j) \in H$ and

$$\langle h^{N_k}, k(\cdot, t_j)\rangle_H = h^{N_k}(t_j) \underset{k\to+\infty}{\rightharpoonup} \langle h^*, k(\cdot, t_j)\rangle_H = h^*(t_j).$$

Therefore, for all $N \geq 1$,

$$\pi_N(h^{N_k}) \underset{k\to+\infty}{\longrightarrow} \pi_N(h^*)$$

in the finite-dimensional space H_N. According to assumption ($H1$) (i.e., $\Delta_N \subset \Delta_{N+1}$), as far as $N_k \geq N$

$$\pi_N(h^{N_k}) = \pi_N(\rho_{N_k}(\widehat{u}_{N_k})) = \pi_N(\widehat{u}_{N_k}),$$

so that

$$\pi_N(\widehat{u}_{N_k}) \underset{k\to+\infty}{\longrightarrow} \pi_N(h^*) \quad \text{in } H_N.$$

As H_N is a Hilbertian subspace of E (Inequality (19) of Lemma 5.1),

$$\pi_N(\widehat{u}_{N_k}) \underset{k\to+\infty}{\longrightarrow} \pi_N(h^*) \quad \text{in } E.$$

Under hypothesis ($H2$), $\pi_N(\widehat{u}_{N_k}) \in C$ and C is closed in E, so that for all N,

$$\pi_N(h^*) \in C.$$

The set C is closed in E and $\pi_N(h^*) \underset{N\to+\infty}{\longrightarrow} h^*$ in E, then

$$h^* \in C \quad \text{and} \quad J(\widehat{u}) \leq J(h^*).$$

Then, as J is convex and lower semi continuous and $h^{N_k} \underset{k\to+\infty}{\rightharpoonup} h^* \in H$, using (31),

$$\begin{aligned} J(\widehat{u}) &\leq J(h^*) \leq \underline{\lim}_k J(h^{N_k}) = \underline{\lim}_k J_{N_k}(\widehat{u}_{N_k}) + \underbrace{c_3\varepsilon_{N_k}}_{=0} \\ &\leq \overline{\lim}_k J_{N_k}(\widehat{u}_{N_k}) + \underbrace{c_3\varepsilon_{N_k}}_{=0} \leq J(\widehat{u}) + \underbrace{\overline{\lim}_k c_1\eta_N + c_3\varepsilon_N}_{=0} \end{aligned}$$

so that

$$J_{N_k}(\widehat{u}_{N_k}) \underset{k\to+\infty}{\longrightarrow} J(\widehat{u}).$$

Equation (31) implies that

$$J_{N_k}(\pi_{N_k}(\widehat{u})) \underset{k\to+\infty}{\longrightarrow} J(\widehat{u}).$$

As the sequences $(J_N(\widehat{u}_N))_{N\in\mathbb{N}}$ and $(J_N(\pi_N(\widehat{u}))_{N\in\mathbb{N}}$ are in the compact set $[0, J(\widehat{u}) + \sup_N(c_1\eta_N + c_3\varepsilon_N)]$, then the results hold. □

References

[1] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68**.

[2] Bay, X., Grammont, L. and Maatouk, H. (2016). Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation. *Electron. J. Statist.* **10** 1580–1595.

[3] Bay, X., Grammont, L. and Maatouk, H. (2017). A new method for interpolating in a convex subset of a Hilbert space. *Computational Optimization and Applications* **68** 95–120.

[4] Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

[5] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

[6] Cai, B. and Dunson, D. B. (2007). Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *Journal of the American Statistical Association* **102** 1158–1171.

[7] Chataigner, M., Cousin, A., Crépey, S., Dixon, M. and Gueye, D. (2021). Beyond surrogate modeling: Learning the local volatility via shape constraints. *SIAM Journal on Financial Mathematics* **12** SC58–SC69.

[8] Cousin, A., Deleplace, A. and Misko, A. (2022). Gaussian Process Regression for Swaption Cube Construction under No-Arbitrage Constraints. *Risks* **10** 232.

[9] Cousin, A., Maatouk, H. and Rullière, D. (2016). Kriging of financial term-structures. *European Journal of Operational Research* **255** 631–648.

[10] Cramer, H. and Leadbetter, R. (1967). *Stationary and related stochastic processes: sample function properties and their applications. Wiley series in probability and mathematical statistics. Tracts on probability and statistics*. Wiley.

[11] Curtis, S. M. and Ghosh, S. K. (2011). A variable selection approach to monotonic regression with Bernstein polynomials. *Journal of Applied Statistics* **38** 961-976.

[12] Golchi, S., Bingham, D. R., Chipman, H. and Campbell, D. A. (2015). Monotone Emulation of Computer Experiments. *SIAM/ASA Journal on Uncertainty Quantification* **3** 370-392.

[13] Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* **27** 1-33.

[14] Kimeldorf, G. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* 495–502.

[15] Lenk, P. J. and Choi, T. (2017). Bayesian analysis of shape-restricted functions using Gaussian process priors. *Statistica Sinica* 43–69.

[16] Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317.

[17] López-Lopera, A. F., Bachoc, F., Durrande, N., Rohmer, J.,

Idier, D. and Roustant, O. (2018). Approximating Gaussian process emulators with linear inequality constraints and noisy observations via MC and MCMC. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* 363–381. Springer.

[18] López-Lopera, A. F., Bachoc, F., Durrande, N. and Roustant, O. (2018). Finite-Dimensional Gaussian Approximation with Linear Inequality Constraints. *SIAM/ASA Journal on Uncertainty Quantification* **6** 1224-1255.

[19] Maatouk, H. (2022). Finite-dimensional approximation of Gaussian processes with linear inequality constraints and noisy observations. *Communications in Statistics - Theory and Methods* **0** 1-20.

[20] Maatouk, H. and Bay, X. (2016). A new rejection sampling method for truncated multivariate Gaussian random variables restricted to convex sets. In *Monte carlo and quasi-monte carlo methods* 521–530. Springer.

[21] Maatouk, H. and Bay, X. (2017). Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences* **49** 557–582.

[22] Maatouk, H., Bay, X. and Rullière, D. (2022). A note on simulating hyperplane-truncated multivariate normal distributions. *Statistics & Probability Letters* **191** 109650.

[23] Meyer, M. C., Hackstadt, A. J. and Hoeting, J. A. (2011). Bayesian estimation and inference for generalised partial linear models using shape-restricted splines. *Journal of Nonparametric Statistics* **23** 867–884.

[24] Pakman, A. and Paninski, L. (2014). Exact hamiltonian monte carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics* **23** 518–542.

[25] Parzen, E. (1959). *Statistical inference on time series by Hilbert space methods*. Stanford University.

[26] Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[27] Ray, P., Pati, D. and Bhattacharya, A. (2020). Efficient Bayesian shape-restricted function estimation with constrained Gaussian process priors. *Statistics and Computing* **30** 839–853.

[28] Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* 645–652. JMLR Workshop and Conference Proceedings.

[29] Schwartz, L. (1964). Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'analyse mathématique* **13** 115–256.

[30] Shively, T. S., Walker, S. G. and Damien, P. (2011). Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *Journal of Econometrics* **161** 166–181.

[31] Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C. and Jakeman, J. D. (2020). A survey of constrained Gaussian process regression:

Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing* **1**.

[32] Zhou, S., Giulani, P., Piekarewicz, J., Bhattacharya, A. and Pati, D. (2019). Reexamining the proton-radius problem using constrained Gaussian processes. *Phys. Rev. C* **99** 055202.

[33] Zhou, S., Ray, P., Pati, D. and Bhattacharya, A. (2022). A mass-shifting phenomenon of truncated multivariate normal priors. *Journal of the American Statistical Association* **0** 1-37.

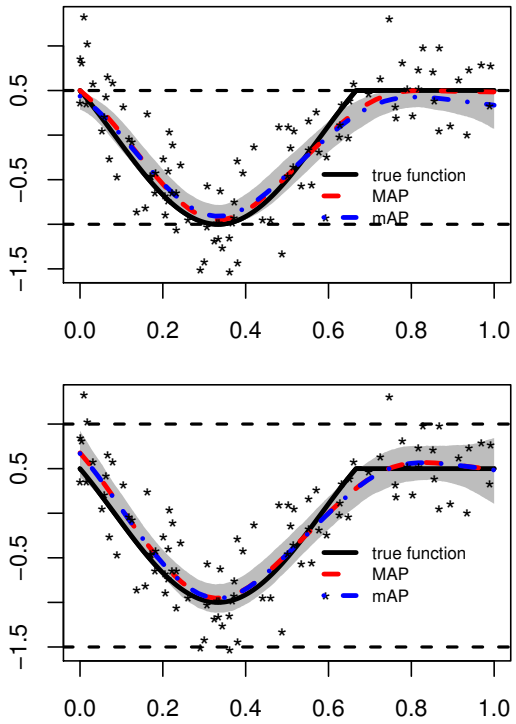

Fig 1: Comparison of estimates for the bounded flat function f_1 using the GP approach with boundedness constraints. The black solid curve represents the true function, while the red dashed curve shows the MAP estimate and the blue dashed-dotted curve shows the mAP estimate. The black stars represent the 100 training data generated from (1) using the true f_1 and a true noise variance $\sigma^2_{\text{noise}} = 0.4^2$. The gray shaded area indicates the 95% pointwise confidence interval. The two dashed horizontal lines represent the lower and upper bounds: $[-1, 0.5]$ on the left and $[-1.5, 1]$ on the right.

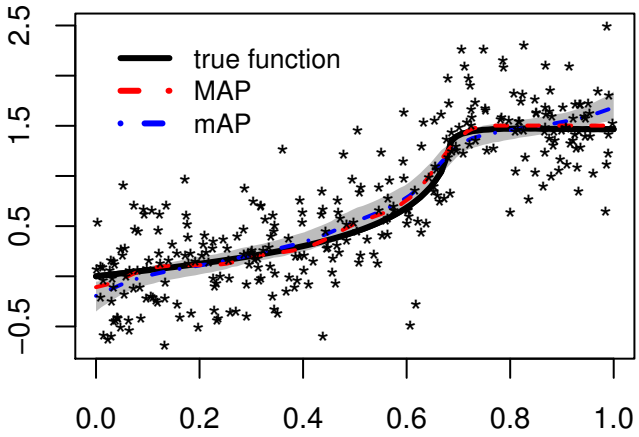

Fig 2: Prediction accuracy for f_2 using the proposed GP approach with monotonicity constraints: the black solid curve corresponds to the true function, the red dashed curve represents the MAP estimate, the blue dashed-dotted curve corresponds to the mAP estimate, and the black stars are the 300 training data generated from (1) with f_2 and a true noise variance $\sigma^2_{\text{noise}} = 0.4^2$.

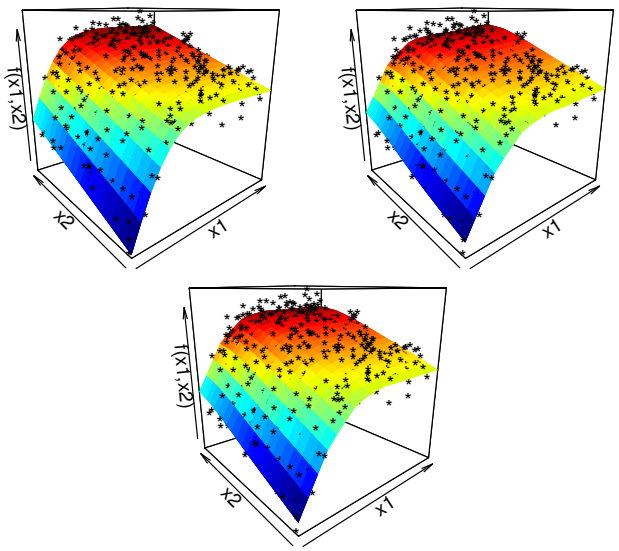

Fig 3: Left: the true nondecreasing function with respect to the two inputs together with the training data (black stars). Middle (respectively, Right): the mAP estimate (respectively, MAP estimate) from the GP approach defined in Equation (30) with the training data.

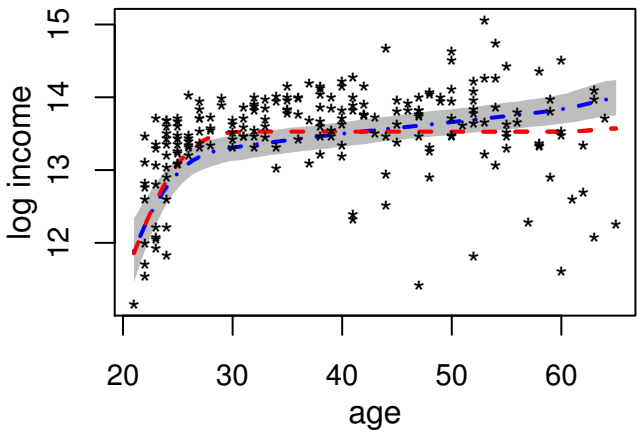

Fig 4: The estimation accuracy of the proposed approach applied to the age-income data is shown. The red dashed curve corresponds to the MAP estimate, while the blue dashed-dotted curve corresponds to the mAP estimate. The black stars represent the noisy data, and the gray shaded area represents the 95% pointwise confidence interval.

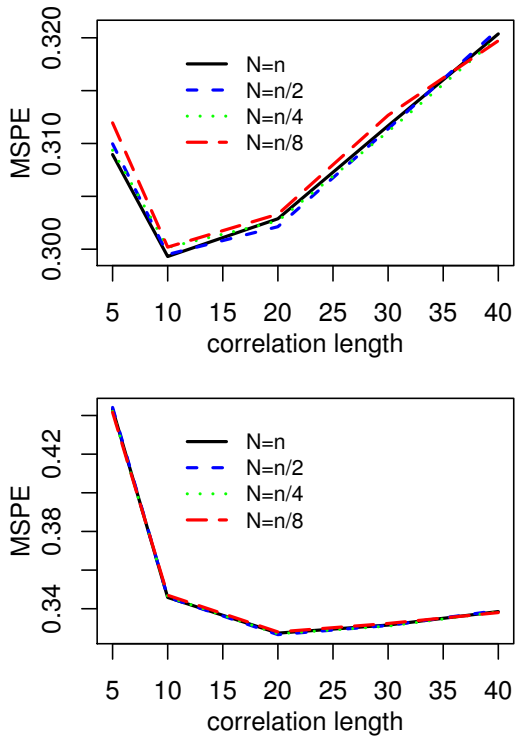

Fig 5: MSPEs as a function of the correlation length parameter, θ, using the 5-fold cross-validation technique for different values of $N \in \{n, \lfloor n/2 \rfloor, \lfloor n/4 \rfloor, \lfloor n/8 \rfloor\}$ and standard deviation, σ_{noise}, equal to 0.5 (left panel) and 1 (right panel).