Weighted Leave-One-Out Cross Validation

Luc Pronzato, Maria-João Rendas

Weighted Leave-One-Out Cross Validation*

Luc Pronzato† Maria-João Rendas†

September 5, 2024

Abstract

We present a weighted version of Leave-One-Out (LOO) cross-validation for estimating the Integrated Squared Error (ISE) when approximating an unknown function by a predictor that depends linearly on evaluations of the function over a finite collection of sites. The method relies on the construction of the best linear estimator of the squared prediction error at an arbitrary unsampled site based on squared LOO residuals, assuming that the function is a realization of a Gaussian Process (GP). A theoretical analysis of performance of the ISE estimator is presented, and robustness with respect to the choice of the GP kernel is investigated first analytically, then through numerical examples. Overall, the estimation of ISE is significantly more precise than with classical, unweighted, LOO cross validation. Application to model selection is briefly considered through examples.

Keywords Leave-one-out cross validation, integrated squared prediction error, computer experiments, space-filling design

MSCcodes 65D05, 62G99, 62-08

1 Introduction

The paper addresses the characterization of the performance of data-driven model learning. We consider the fairly general setting where a learning dataset collecting the evaluations of an unknown function f at a given set of sites $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is used to predict the value of f at generic points $\mathbf{x}$ in some set $\mathscr{X}$. When function evaluations are computationally expensive (for example when they involve complex computer simulations) their number is necessarily limited and the selection of appropriate sites $\mathbf{x}_i$ is crucial, a problem addressed by the experimental design literature. Regardless of the sites chosen and the prediction method used, it is important to assess the quality of the predictions produced by the learned model. The most common measure of performance is the Integrated Squared (prediction) Error (ISE) for a given measure of interest μ over $\mathscr{X}$, and this is the criterion considered in this paper.

Gaussian Process (GP) models are commonly used in computer experiments to formalize prior knowledge about the behavior of the unknown f, as they provide access to the full Bayesian machinery: considering f as a realization of a GP, it is possible to encode prior knowledge on f, such as its regularity over $\mathscr{X}$, and function evaluations $f(\mathbf{x}_i)$, considered as observations $y_i = Y_{\mathbf{x}_i}$ on a sample path of the GP, can be used to update knowledge about f in a Bayesian

*This work was partially funded by project ANR INDEX (ANR-18-CE91-0007).

†CNRS, Université Côte d'Azur, Laboratoire I3S, Sophia Antipolis, France (luc.pronzato@univ-cotedazur.fr, https://www.i3s.unice.fr/~pronzato/, rendas@univ-cotedazur.fr).

way. Although in computer experiments GP models are traditionally used to predict values of f, in this work we use them to infer the performance of a given predictor, by exploiting the property that fourth-order moments of Gaussian variables are directly available. The predictor whose performance is inferred may be itself the Best Linear Unbiased Predictor (BLUP) derived from a GP model, but this is not mandatory and our approach applies to any predictor *linear* in the observations y_i whose weights may depend arbitrarily on $\mathbf{x}$. Linearity in the observations is essential in order to preserve the Gaussianity of prediction errors, but this covers a wide range of prediction methods (extension to non-linear predictors is theoretically possible, through Taylor series expansion and the calculation of higher-order moments of Gaussian variables, but seems tedious and rather unpractical). Also, the paper focuses on the case of noise-free observations, i.e. situations where we directly observe the response of a deterministic simulator, but the presence of observational noise can be taken into account by introducing a nugget effect into the GP model, and the modifications this implies are briefly indicated in Section 4.7 (some simulation results for noisy observations are presented in Section D.2).

A classical approach to evaluate the performance of a prediction method without using observations other than those used to learn the prediction model itself is Cross Validation (CV), and in particular Leave-One-Out Cross Validation (LOOCV). In LOOCV, the value of f at each site $\mathbf{x}_i$ is predicted by the same method but removing $\mathbf{x}_i$ from the learning dataset; the difference with the observed value y_i determines a residual error ε_{-i}, and the empirical average of the ε_{-i}^2 is used to quantify the overall quality of the prediction method. A large number of papers have addressed cross validation, in particular questioning in what sense it can be considered as an estimate of a statistic of the prediction error, see e.g. [4]. In general, the performance analysis considers that LOOCV provides an estimate of the expected squared prediction error, assuming that the learning dataset is an independent and identically distributed (i.i.d.) sample from the joint distribution of the model covariates and outputs. Additionally, it usually assumes that the available observations are noisy versions of the model output, an assumption that we relax here. We will show that, as it can be anticipated, for a well spread design $\mathbf{X}_n$ LOOCV tends to strongly overestimate the actual conditional ISE, given the learning dataset. The estimator we propose, by using the geometry of the design on which the model has been learned to weight the LOO residuals, is able to overcome this drawback. As an unsought feature, our method is able to cope with the problem of covariate shift, of current interest amongst the community of machine learning [23, 24] and to which CV is known to be highly sensitive. Covariate shift occurs when the design in the learning dataset is not a sample from the target distribution under which we want to assess the prediction error: the weights of the corrected LOOCV estimator that we propose depend on the design used to learn the model studied, and thus automatically adjust to covariate shift.

The main objective of the paper is to propose an estimate of the ISE that overcomes some of the limitations of the LOOCV methodology by relying on a GP model, not necessarily stationary, for the fitted function f (the possible extension to mixtures of GP models is briefly considered in Section F in the supplement). This allows us to estimate the expected squared prediction error, conditioned on the learning dataset, the expectation being now taken with respect to f. Our method builds on [18] and relies on the construction of the best linear estimate of the squared prediction error at any $\mathbf{x} \in \mathscr{X}$ based on the set of squared LOO residuals ε_{-i}^2. Integration with respect to $\mathbf{x}$ (a simple summation when the measure μ is discrete) then provides the ISE estimate. The numerical experiments presented show that our method significantly improves the accuracy of performance evaluation compared to the straightforward application of LOOCV.

As the resulting performance measure depends on the assumed GP model, the paper and its supplement include extensive numerical analyses that confirm robustness with respect to its choice.

2 Notation, motivation and paper organization

2.1 Notation

Consider an n-point design $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ without repetitions, i.e., $\mathbf{x}_i \neq \mathbf{x}_j$ for all $i \neq j$, and let $\mathbf{y}_n = [y_1, \ldots, y_n]^\top$, where $y_i = f(\mathbf{x}_i)$ is the observation at site $\mathbf{x}_i$[1]. We denote by $\mathscr{F}_n$ the learning dataset: $\mathscr{F}_n = \{(\mathbf{x}_i, y_i)_{i=1}^n\}$ — a sub-index n will indicate dependency of the corresponding quantity on $\mathscr{F}_n$ (sometimes, only on the design $\mathbf{X}_n$).

Let $\eta_n(\mathbf{x})$ be the (arbitrary) predictor of $f(\mathbf{x})$ at an unsampled site $\mathbf{x}$, learned using $\mathscr{F}_n$, whose performance we want to assess, and denote by

$$\varepsilon_n(\mathbf{x}) = f(\mathbf{x}) - \eta_n(\mathbf{x})$$

its prediction error at $\mathbf{x}$. Our goal is to estimate, using only the dataset $\mathscr{F}_n$, the ISE for some given positive measure of importance μ over $\mathscr{X}$:

$$\mathsf{ISE}(\eta_n) = \int_{\mathscr{X}} \varepsilon_n^2(\mathbf{x})\, \mu(d\mathbf{x})\,. \tag{2.1}$$

Without any loss of generality, we assume that $\mathbf{X}_n \in \mathscr{X}^n$. All predictors η_n considered in the paper are linear, being defined by a weight function $\mathbf{w}(\cdot,\cdot) : (\mathbf{x}, \mathbf{X}_n) \in \mathscr{X} \times \mathscr{X}^n \to \mathbf{w}(\mathbf{x}, \mathbf{X}_n) \in \mathbb{R}^n$, such that the prediction of $f(\mathbf{x})$ based on $\mathbf{y}_n$ is $\eta_n(\mathbf{x}) = \mathbf{w}^\top(\mathbf{x}, \mathbf{X}_n)\mathbf{y}_n$, for any $n \in \mathbb{N}$ and any $\mathbf{X}_n$. In the following, we use the simpler notation $\mathbf{w}_n(\mathbf{x}) = \mathbf{w}(\mathbf{x}, \mathbf{X}_n)$ — note that η_n does not necessarily interpolate the data, i.e., we may have $\eta_n(\mathbf{x}_i) \neq y_i$; the prediction error is therefore not necessarily null at the $\mathbf{x}_i$. We always assume that $\mathbf{w}_n(\cdot)$ is bounded on $\mathscr{X}$.

The estimator of $\mathsf{ISE}(\eta_n)$ proposed in this paper, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, assumes that f is a realization of a GP $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$, indexed by $\mathbf{x}$ in $\mathscr{X}$. Here K is a Strictly Positive Definite (SPD) kernel and $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$ means that $\mathsf{E}\{Y_{\mathbf{x}}\} = 0$ and $\mathsf{E}\{Y_{\mathbf{x}} Y_{\mathbf{x}'}\} = \sigma^2 K(\mathbf{x}, \mathbf{x}')$ for all $\mathbf{x}$ and $\mathbf{x}'$ in $\mathscr{X}$. Throughout the paper, $\mathbf{k}_n(\mathbf{x})$ is the vector with components $K(\mathbf{x}, \mathbf{x}_i)$, $i = 1, \ldots, n$, and $\mathbf{K}_n$ is the $n \times n$ matrix with $\{\mathbf{K}_n\}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, n$. We will denote $\mathbf{M}_n = \mathbf{K}_n^{-1}$ (and to simplify notation the subindex n will be dropped, i.e., $\mathbf{M} \equiv \mathbf{M}_n$).

The assumption of a GP model for f is important for at least three reasons: (i) it is extremely convenient for evaluating the performance of an arbitrary ISE estimator for a given linear predictor η_n; (ii) it is essential for the derivation of the ISE estimator we propose; (iii) it is commonly used to define a predictor η_n. We thus need to carefully distinguish between three possibly different GP models in the developments presented below.

We will denote by $\mathsf{GP}(0, \sigma^2 K)$ the ("true") data generating model. In a real situation, the data $f(\mathbf{x}_i)$ are not samples from $\mathsf{GP}(0, \sigma^2 K)$, but this assumption allows explicit computation of the bias and Mean Squared Error (MSE) of the different ISE estimators considered. When we will need to distinguish it from $\mathsf{GP}(0, \sigma^2 K)$, the GP model assumed for the construction of our estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ will be denoted by $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$. Indeed, there is no reason to assume that $\mathsf{GP}(0, \sigma^2 K) \equiv \mathsf{GP}(0, \sigma_e^2 K^{(e)})$, and we will investigate the impact of the possible modeling

[1] By default all vectors are column vectors.

mismatch $\mathsf{GP}(0, \sigma^2 K) \neq \mathsf{GP}(0, \sigma_e^2 K^{(e)})$. Finally, the predictor η_n for which we want to estimate the ISE may itself rely on a certain GP model, which we shall denote by $\mathsf{GP}(0, \sigma_p^2 K^{(p)})$, where $K^{(p)}$ differs in general from K and $K^{(e)}$. When necessary, a superscript $^{(e)}$ or $^{(p)}$ will explicitly indicate the underlying model to which we refer. An essential feature of our method is that it relies solely on predictions (it does not require estimation of the scaling parameter σ^2), so the choice of kernel is not very critical: indeed, predictions based on GP models are known to be robust to the choice of kernel K, which is not the case for the prediction of their accuracy; see, e.g., [21, Sect. 3.5]; one may also refer to [1, 14, 16] for results on the estimation of σ^2 in a GP model and consequences on model calibration.

That we always suppose a GP with zero mean may appear as a severe restriction, as the hypothesis according to which f is a realization of a centred GP is often difficult to maintain in practice. However, the presence of a linear trend $\boldsymbol{\tau}^\top \mathbf{h}(\mathbf{x})$, with $\mathbf{h}(\cdot) = [h_1(\cdot), \ldots, h_n(\cdot)]^\top$ a vector of p functions on $\mathscr{X}$, has no effect on our ISE estimator when the predictor satisfies $[h_i(\mathbf{x}_1), \ldots, h_i(\mathbf{x}_n)]\mathbf{w}_n(\mathbf{x}) = h_i(\mathbf{x})$ for all $i = 1, \ldots, p$, all n and all $\mathbf{x}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This is the case in particular for universal kriging (and when the trend is a constant mean, i.e., $\mathbf{h}(\mathbf{x}) \equiv 1$ for all $\mathbf{x}$, the condition $\sum_{i=1}^n w_i(\mathbf{x}) = 1$ for all n and all $\mathbf{x}$ is satisfied by the ordinary kriging predictor). See Section E in the supplement for details, including a simple and convenient modification of the estimator for the case when the weights $\mathbf{w}_n(\mathbf{x})$ do not satisfy the condition above.

The kernels used in the examples presented in the paper are all isotropic (i.e., $K(\mathbf{x}, \mathbf{x}')$ only depends on $\|\mathbf{x} - \mathbf{x}'\|$). However, this is not mandatory and any kernel $K^{(e)}$ can be used to construct $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$. This means that, independently of the particular predictor whose performance is to be assessed, we may choose the kernel we think is the most appropriate as a GP model for f: $K^{(e)}$ can be non isotropic, non stationary, and one may even consider a mixture of GP models (see Section F in the supplement), which offers considerable flexibility.

2.2 Motivation

To motivate our work and precise the basic notions that we shall use, in this introductory section we consider the case when $\mathsf{GP}(0, \sigma_e^2 K^{(e)}) = \mathsf{GP}(0, \sigma_p^2 K^{(p)}) = \mathsf{GP}(0, \sigma^2 K)$ and $\eta_n \equiv \eta_n^*$, the Best Linear Unbiased Predictor (BLUP) of f given $\mathscr{F}_n$; that is, η_n is the simple kriging interpolator (the posterior expectation under the GP model):

$$\eta_n^*(\mathbf{x}) = \mathsf{E}\{Y_\mathbf{x} | \mathscr{F}_n\} = \mathbf{k}_n^\top(\mathbf{x}) \mathbf{K}_n^{-1} \mathbf{y}_n \,. \tag{2.2}$$

The posterior variance of $Y_\mathbf{x}$ is independent of the observations $\mathbf{y}_n$: $\mathsf{var}\{Y_\mathbf{x} | \mathscr{F}_n\} = \mathsf{E}\{[Y_\mathbf{x} - \eta_n^*(\mathbf{x})]^2 | \mathscr{F}_n\} = \mathsf{E}\{[Y_\mathbf{x} - \eta_n^*(\mathbf{x})]^2 | \mathbf{X}_n\}$, and direct calculation gives $\mathsf{var}\{Y_\mathbf{x} | \mathscr{F}_n\} = \sigma^2 \, {\rho_n^*}^2(\mathbf{x})$, where

$${\rho_n^*}^2(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) - \mathbf{k}_n^\top(\mathbf{x}) \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x}) \tag{2.3}$$

is the (simple) kriging variance. As the learning design $\mathbf{X}_n$ is fixed, all expectations are conditioned on $\mathbf{X}_n$ but in the following we shall simply write $\mathsf{E}\{\cdot\} = \mathsf{E}\{\cdot | \mathbf{X}_n\}$. The assumption that $\eta_n \equiv \eta_n^*$ for a given GP model will be flagged with an asterisk.

For a given (positive) measure of importance μ on $\mathscr{X}$, the Integrated Squared Error $\mathsf{ISE}(\eta_n)$ defined in (2.1) is a natural criterion for measuring the overall predictive quality of η_n over $\mathscr{X}$. Since $Y_\mathbf{x}$, and thus $\varepsilon_n(\mathbf{x})$, is unknown for $\mathbf{x} \notin \mathbf{X}_n$, $\mathsf{ISE}(\eta_n)$ is not computable. However, under

the GP assumption for $Y_{\mathbf{x}}$ its expected value (the Integrated Mean Squared Error) is given by

$$\mathsf{IMSE}(\eta_n^*) = \int_{\mathscr{X}} \mathsf{E}\{\varepsilon_n^2(\mathbf{x})|\mathscr{F}_n\}\,\mu(\mathrm{d}\mathbf{x}) = \sigma^2 \int_{\mathscr{X}} {\rho_n^*}^2(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x})\,. \tag{2.4}$$

As the process variance σ^2 only appears as a multiplicative factor in (2.4), the design $\mathbf{X}_n$ that minimizes $\mathsf{IMSE}(\eta_n^*)$ is independent of σ^2; see e.g. [20] for an early reference and [9, 10] for methods that avoid repeated computations of integrals when constructing an IMSE-optimal design. Note that unlike the method proposed in this paper, which further exploits knowledge of $\mathscr{F}_n$, direct use of (2.4) to quantify the predictive quality of the model learned over a design $\mathbf{X}_n$ requires knowledge of σ^2, the process variance[2].

Several methods have been proposed to assess the performance of predictors using the learning dataset $\mathscr{F}_n$, amongst which cross validation and in particular the LOOCV criterion [22]:

$$\widehat{\mathsf{ISE}}_{LOO}(\eta_n) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{-i}^2\,, \tag{2.5}$$

where $\varepsilon_{-i} = y_i - \eta_{n\setminus i}(\mathbf{x}_i)$, with $\eta_{n\setminus i}$ the predictor constructed without the i-th design point $\mathbf{x}_i$. As explained in Section 3.1, the actual performance of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ as an estimator of $\mathsf{ISE}(\eta_n)$ strongly depends on the design used to construct η_n. The ISE estimator that we propose optimally weights the squared LOO residuals ε_{-i}^2 by constructing the Best Linear Predictor (BLP) of $\varepsilon_n^2(\mathbf{x})$ at an unsampled site $\mathbf{x}$ (it minimizes a squared loss and is linear in the ε_{-i}^2), and thereby ensures robust performance with respect to the design configuration. Central to the method is the fact that all moments required to calculate this BLP are directly available, thanks to the Gaussian assumption.

2.3 Paper organization

In Section 3.1, we derive the bias, variance and MSE of ISE estimators that are linear in the squared LOO residuals ε_{-i}^2, under the GP model assumption and for a general linear predictor η_n. Results for $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ follow as a particular case. The special case where $\eta_n \equiv \eta_n^*$, the BLUP for the assumed GP model, is considered in Section 3.2. In Section 4.1, we derive our estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, the integral of the BLP $\widehat{\varepsilon_n^2}(\mathbf{x})$ of the squared prediction error $\varepsilon_n^2(\mathbf{x})$ based on the squared LOO residuals ε_{-i}^2. Its mean and MSE are given in Section 4.2 under the assumption that $K^{(e)} = K$ (i.e., in absence of modeling error). In Section 4.3 we assume that $\eta_n \equiv \eta_n^*$, the BLUP for $\mathsf{GP}(0, \sigma^2 K)$, and in Section 4.4 we consider the performance of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ for a general linear predictor η_n in presence of model misspecification ($K^{(e)} \neq K$). Two limiting behaviors for $K^{(e)}$ are briefly investigated in Section 4.5: the independent limit where the correlation between $Y_{\mathbf{x}}$ and $Y_{\mathbf{x}'}$ is negligible when $\mathbf{x} \neq \mathbf{x}'$, the flat limit where conversely $Y_{\mathbf{x}}$ and $Y_{\mathbf{x}'}$ remain strongly correlated when $\|\mathbf{x}' - \mathbf{x}\|$ is large. The BLP $\widehat{\varepsilon_n^2}(\mathbf{x})$ of Section 4.1 is biased, and therefore $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is biased too; a bias-corrected version (in absence of model misspecification) is presented in Section 4.6. Modifications implied by the presence of observation noise are briefly discussed in Section 4.7. A numerical investigation of the performance of the estimators

[2]Numerical investigations show that methods that rely on the estimation of σ^2, e.g. by Maximum Likelihood or LOOCV, and estimate $\mathsf{ISE}(\eta_n)$ by $\mathsf{IMSE}(\eta_n)$ perform worse than the one proposed in the paper; see Section 6 for a brief discussion.

$\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is carried out in Section 5 where several numerical examples are presented. First, in Section 5.1 we illustrate the influence of the design $\mathbf{X}_n$ on both estimators for a univariate function f. Then, in Section 5.2 we study the robustness of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ with respect to the assumed kernel $K^{(e)}$, considering kernels with different correlation lengths; different predictors η_n are also considered: a non-interpolating polynomial (Section 5.2.1) and the BLUP for a GP model $\mathsf{GP}(0, \sigma_p^2 K^{(p)})$ (Section 5.2.2). Additional material is provided in Section B of the supplement and involves kernels $K^{(e)}$ with different regularities. While in Sections 5.1 and 5.2 $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$, in Sections 5.3 and 5.4 we consider test-cases from the literature, with f depending on 2 and 4 variables, respectively. Other numerical results are presented in the supplement: average performance characteristics of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ for GP realizations with $n \in \{10\,d, 20\,d, 50\,d, 100\,d, 200\,d\}$ and $d \in \{4, 6, 8\}$ in Section C; empirical performance for random functions that are not GP realizations in Section D, including the introduction of observation noise in Section D.2. Section 6 concludes and draws some perspectives for this work.

That we propose to estimate the ISE (2.1) by another integral may seem numerically cumbersome. However, the integrated function is known explicitly, and the integral can be approximated by Quasi-Monte Carlo with small computational cost — an indication of computational times for a Matlab implementation is provided in Sections 5.3, 5.4 and D.1 in the supplement. In fact, the method proposed does not address the integration problem itself, but rather the estimation $\widehat{\varepsilon_n^2}(\mathbf{x})$, in the mean square sense, of the squared prediction error $\varepsilon_n^2(\mathbf{x})$ at any $\mathbf{x}$. Note, however, that we cannot predict precisely the value of $\varepsilon_n^2(\mathbf{x})$ at every $\mathbf{x}$ (and an example presented in Section D.3 shows that the approach is unreliable for estimating a tail characteristic such as a quantile or a conditional value-at-risk of $\widehat{\varepsilon_n^2}(X)$ with $X \sim \mu$, for one realization of $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2\, K)$, as the estimated value $\widehat{\varepsilon_n^2}(X)$ and $\varepsilon_n^2(X)$ have different distributions). Estimation of $\mathsf{ISE}(\eta_n)$ by the integral of $\widehat{\varepsilon_n^2}(\mathbf{x})$ can be justified by ergodicity arguments when K is stationary, $K(\mathbf{x}, \mathbf{x}') \to 0$ as $\|\mathbf{x} - \mathbf{x}'\| \to \infty$, and the support of μ is large enough.

3 LOOCV for linear predictors under the GP assumption

For linear predictors, of the form $\eta(\mathbf{x}) = \mathbf{w}_n^\top(\mathbf{x})\mathbf{y}_n$, the vector of LOO residuals $\boldsymbol{\varepsilon}_{LOO} = (\varepsilon_{-1}, \ldots, \varepsilon_{-n})^\top$ is also linear in $\mathbf{y}_n$ and can be written as

$$\boldsymbol{\varepsilon}_{LOO} = \mathbf{R}_n^\top \mathbf{y}_n\,, \tag{3.1}$$

where $\mathbf{R}_n = \mathbf{I}_n - \mathbf{W}_{n\backslash}$, with $\mathbf{I}_n$ the n-dimensional identity matrix and $\mathbf{W}_{n\backslash}$ an $n \times n$ matrix whose diagonal is identically null (as $\varepsilon_{-i} = y_i - \eta_{n\backslash i}(\mathbf{x}_i)$ and $\eta_{n\backslash i}(\mathbf{x}_i)$ does not depend on y_i). We assume that $\mathbf{R}_n$ has full rank.

The assumption that f is a realization of a GP $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$ allows us to derive the expressions of the first two moments of an arbitrary linear combination $\boldsymbol{\gamma}^\top \boldsymbol{\varepsilon}_{LOO}^{\odot 2}$ of the squared LOO residuals $\boldsymbol{\varepsilon}_{LOO}^{\odot 2} = (\varepsilon_{-1}^2, \ldots, \varepsilon_{-n}^2)^\top$, $\boldsymbol{\gamma} \in \mathbb{R}^n$. (Here and in the following we denote by $\mathbf{A}^{\odot 2}$ the Hadamard square of matrix $\mathbf{A}$: $\{\mathbf{A}^{\odot 2}\}_{ij} = \mathbf{A}_{ij}^2$.) As the LOOCV criterion $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ corresponds to $\boldsymbol{\gamma} = \mathbf{1}_n/n$ with $\mathbf{1}_n$ the n-dimensional vector with all components equal to 1, see (2.5), we directly obtain the bias and MSE of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$.

Table 1: Moments of quantities of interest.

	b	σ^2	$\boldsymbol{\varepsilon}_{LOO}^{\odot 2\top}$	$\sigma\,\varepsilon_n(\mathbf{x})$	$\varepsilon_n^2(\mathbf{x})$	$\mathsf{ISE}(\eta_n)$
a	$\mathsf{E}\{ab\}/\sigma^4$	$*$	$*$	$*$	$*$	$*$
σ^2	$*$	1	$\mathbf{u}_n^\top$	0	$\rho_n^2(\mathbf{x})$	J_n
$\boldsymbol{\varepsilon}_{LOO}^{\odot 2}$	$*$	$\mathbf{u}_n$	$\mathbf{S}_n$	$\mathbf{0}$	$\mathbf{c}_n(\mathbf{x})$	$\mathbf{b}_n$
$\sigma\,\varepsilon_n(\mathbf{x}')$	$*$	0	$\mathbf{0}$	$\rho_n^2(\mathbf{x},\mathbf{x}')$	0	0
$\varepsilon_n^2(\mathbf{x}')$	$*$	$\rho_n^2(\mathbf{x}')$	$\mathbf{c}_n^\top(\mathbf{x}')$	0	$\rho_n^2(\mathbf{x})\rho_n^2(\mathbf{x}')+2\,\rho_n^4(\mathbf{x},\mathbf{x}')$	
$\mathsf{ISE}(\eta_n)$	$*$	J_n	$\mathbf{b}_n^\top$	0		$J_n^2+2\,V_n$

3.1 Consequences of the GP assumption and notation

Under the GP assumption $Y_{\mathbf{x}} \sim \mathsf{GP}(0,\sigma^2 K)$, the prediction error $\varepsilon_n(\mathbf{x}) = Y_{\mathbf{x}} - \eta_n(\mathbf{x})$ has zero mean, $\mathsf{E}\{\varepsilon_n(\mathbf{x})\} = 0$, and we denote

$$\rho_n^2(\mathbf{x}) = \mathsf{E}\{\varepsilon_n^2(\mathbf{x})/\sigma^2\} = K(\mathbf{x},\mathbf{x}) - 2\,\mathbf{w}_n^\top(\mathbf{x})\mathbf{k}_n(\mathbf{x}) + \mathbf{w}_n^\top(\mathbf{x})\mathbf{K}_n\mathbf{w}_n(\mathbf{x})\,. \tag{3.2}$$

The vector of LOO residuals $\boldsymbol{\varepsilon}_{LOO}$ is also Gaussian and, for all $\mathbf{x}\in\mathscr{X}$,

$$\mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}\} = \mathbf{0}, \qquad \mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}\boldsymbol{\varepsilon}_{LOO}^\top\} = \sigma^2\mathbf{R}_n^\top\mathbf{K}_n\mathbf{R}_n, \qquad \mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}\boldsymbol{\varepsilon}_n(\mathbf{x})\} = \sigma^2\,\mathbf{R}_n^\top\mathbf{t}_n(\mathbf{x})\,,$$

where

$$\mathbf{t}_n(\mathbf{x}) = \mathsf{E}\{\mathbf{y}_n\varepsilon_n(\mathbf{x})\}/\sigma^2 = \mathbf{k}_n(\mathbf{x}) - \mathbf{K}_n\mathbf{w}_n(\mathbf{x})\,. \tag{3.3}$$

Using the formula for the expectation of the product of squared Gaussian variables a and b,

$$\mathsf{E}\{a^2b^2\} = \mathsf{E}\{a^2\}\mathsf{E}\{b^2\} + 2\,[\mathsf{E}\{ab\}]^2\,,$$

we get $\mathsf{E}\{\varepsilon_n^2(\mathbf{x})\varepsilon_n^2(\mathbf{x}')\} = \sigma^4\left[\rho_n^2(\mathbf{x})\rho_n^2(\mathbf{x}') + 2\,\rho_n^4(\mathbf{x},\mathbf{x}')\right]$, where $\rho_n^2(\mathbf{x})$ is given by (3.2) and

$$\rho_n^2(\mathbf{x},\mathbf{x}') = \frac{1}{\sigma^2}\,\mathsf{E}\left\{\varepsilon_n(\mathbf{x})\varepsilon_n(\mathbf{x}')\right\} = K(\mathbf{x},\mathbf{x}') - \mathbf{w}_n^\top(\mathbf{x})\mathbf{k}_n(\mathbf{x}') - \mathbf{w}_n^\top(\mathbf{x}')\mathbf{k}_n(\mathbf{x}) + \mathbf{w}_n^\top(\mathbf{x})\mathbf{K}_n\mathbf{w}_n(\mathbf{x}').$$

In the same manner, we obtain for the (normalized) first two moments of $\boldsymbol{\varepsilon}_{LOO}^{\odot 2}$,

$$\begin{aligned}
\mathbf{u}_n = \mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}^{\odot 2}\}/\sigma^2 &= \operatorname{diag}\left\{(\mathbf{R}_n^\top\mathbf{K}_n\mathbf{R}_n)\right\}, &(3.4)\\
\mathbf{S}_n = \mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}^{\odot 2}\boldsymbol{\varepsilon}_{LOO}^{\odot 2\top}\}/\sigma^4 &= \mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}^{\odot 2}/\sigma^2\}\mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}^{\odot 2\top}/\sigma^2\} + 2\left(\mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}\boldsymbol{\varepsilon}_{LOO}^\top/\sigma^2\}\right)^{\odot 2}\\
&= \mathbf{u}_n\mathbf{u}_n^\top + 2\,(\mathbf{R}_n^\top\mathbf{K}_n\mathbf{R}_n)^{\odot 2}\,, &(3.5)
\end{aligned}$$

together with

$$\begin{aligned}
\mathbf{c}_n(\mathbf{x}) = \mathsf{E}\{\varepsilon_n^2(\mathbf{x})\boldsymbol{\varepsilon}_{LOO}^{\odot 2}\}/\sigma^4 &= \mathsf{E}\{\varepsilon_n^2(\mathbf{x})/\sigma^2\}\mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}^{\odot 2}/\sigma^2\} + 2\left(\mathsf{E}\{\varepsilon(\mathbf{x})\boldsymbol{\varepsilon}_{LOO}/\sigma^2\}\right)^{\odot 2}\\
&= \rho_n^2(\mathbf{x})\mathbf{u}_n + 2\,[\mathbf{R}_n^\top\mathbf{t}_n(\mathbf{x})]^{\odot 2}\,. &(3.6)
\end{aligned}$$

With these definitions (summarized in Table 1), we can compute the bias, variance and MSE of any linear combination $\widehat{\mathsf{ISE}}(\eta_n) = \boldsymbol{\gamma}^\top \boldsymbol{\varepsilon}_{LOO}^{\odot 2}$, $\boldsymbol{\gamma} \in \mathbb{R}^n$:

$$\begin{aligned}
\mathsf{Bias}\{\widehat{\mathsf{ISE}}(\eta_n)\} &= \mathsf{E}\left\{\widehat{\mathsf{ISE}}(\eta_n) - \mathsf{ISE}(\eta_n)\right\} = \sigma^2\,\boldsymbol{\gamma}^\top \mathbf{u}_n - \mathsf{E}\left\{\mathsf{ISE}(\eta_n)\right\}\,, && (3.7)\\
\mathsf{var}\{\widehat{\mathsf{ISE}}(\eta_n)\} &= \mathsf{E}\left\{\widehat{\mathsf{ISE}}^2(\eta_n)\right\} - \left[\mathsf{E}\left\{\widehat{\mathsf{ISE}}(\eta_n)\right\}\right]^2 = \sigma^4 \boldsymbol{\gamma}^\top (\mathbf{S}_n - \mathbf{u}_n \mathbf{u}_n^\top)\boldsymbol{\gamma} \\
&\qquad = 2\,\sigma^4 \boldsymbol{\gamma}^\top (\mathbf{R}_n^\top \mathbf{K}_n \mathbf{R}_n)^{\odot 2} \boldsymbol{\gamma}\,, && (3.8)\\
\mathsf{MSE}\{\widehat{\mathsf{ISE}}(\eta_n)\} &= \mathsf{E}\left\{\left(\widehat{\mathsf{ISE}}(\eta_n) - \mathsf{ISE}(\eta_n)\right)^2\right\} \\
&= \sigma^4 \boldsymbol{\gamma}^\top \mathbf{S}_n \boldsymbol{\gamma} - 2\,\sigma^4 \boldsymbol{\gamma}^\top \mathbf{b}_n + \mathsf{var}\left\{\mathsf{ISE}(\eta_n)\right\} + \left[\mathsf{E}\left\{\mathsf{ISE}(\eta_n)\right\}\right]^2\,, && (3.9)
\end{aligned}$$

with

$$\begin{aligned}
\mathsf{E}\left\{\mathsf{ISE}(\eta_n)\right\} &= \mathsf{IMSE}(\eta_n) = \sigma^2 J_n\,, && (3.10)\\
\mathsf{var}\left\{\mathsf{ISE}(\eta_n)\right\} &= 2\,\sigma^4\, V_n\,, && (3.11)
\end{aligned}$$

where we have introduced

$$\begin{aligned}
\mathbf{b}_n &= \int_{\mathscr{X}} \mathbf{c}_n(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x})\,, && (3.12)\\
J_n &= \int_{\mathscr{X}} \rho_n^2(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x})\,, && (3.13)\\
V_n &= \int_{\mathscr{X}^2} \rho_n^4(\mathbf{x},\mathbf{x}')\mu(\mathrm{d}\mathbf{x})\mu(\mathrm{d}\mathbf{x}')\,. && (3.14)
\end{aligned}$$

We thus obtain for the LOOCV estimator (2.5) (for which $\boldsymbol{\gamma} = \mathbf{1}_n/n$)

$$\begin{aligned}
\mathsf{Bias}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\} &= \frac{\sigma^2}{n}\mathbf{1}_n^\top \mathbf{u}_n - \mathsf{IMSE}(\eta_n)\,, && (3.15)\\
\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\} &= \sigma^4\left[\frac{\mathbf{1}_n^\top \mathbf{S}_n \mathbf{1}_n}{n^2} - 2\,\frac{\mathbf{1}_n^\top \mathbf{b}_n}{n} + J_n^2 + 2\,V_n\right]\,. && (3.16)
\end{aligned}$$

3.2 BLUP for the assumed GP model

We consider here the special case where η_n is the BLUP η_n^* for the GP from which f is drawn, i.e., $\mathsf{GP}(0, \sigma_p^2 K^{(p)}) \equiv \mathsf{GP}(0, \sigma^2 K)$, a situation where simpler expressions can be found since $\mathbf{w}_n(\mathbf{x}) = \mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})$, see (2.2). Indeed, it implies $\mathbf{t}_n(\cdot) \equiv 0$, see (3.3), and $\rho_n^2(\mathbf{x})$ defined by (3.2) equals ${\rho_n^*}^2(\mathbf{x})$ given by (2.3).

Since the LOO residual ε_{-i} is computed using the BLUP η_n^* that leaves out $(\mathbf{x}_i, y_i)$, i.e.,

$$\varepsilon_{-i} = y_i - \eta_{n\setminus i}^*(\mathbf{x}_i), \qquad i = 1, \ldots, n\,,$$

where $\eta_{n\setminus i}^*(\mathbf{x}) = \mathbf{k}_{n\setminus i}^\top(\mathbf{x})\mathbf{K}_{n\setminus i}^{-1}\mathbf{y}_{n\setminus i}$ (with obvious notation), straightforward use of the block-matrix inversion formula gives

$$\mathbf{R}_n = \mathbf{M}\mathbf{D}_n\,,$$

where $\mathbf{M} = \mathbf{K}_n^{-1}$ and $\mathbf{D}_n = \mathrm{diag}\{1/\mathbf{M}_{ii}, i = 1, \ldots, n\}$, with

$$\mathbf{M}_{ii} = \left[K(\mathbf{x}_i, \mathbf{x}_i) - \mathbf{k}_{n\setminus i}^\top(\mathbf{x}_i)\mathbf{K}_{n\setminus i}^{-1}\mathbf{k}_{n\setminus i}^\top(\mathbf{x}_i)\right]^{-1} = 1/[{\rho_{n\setminus i}^*}^2(\mathbf{x}_i)]$$

the i-th diagonal element of $\mathbf{M}$; see [7] (one may also refer to [11] for the extension to multiple-fold CV). Note that in this case $\mathbf{R}_n$ always has full rank.

To highlight the difference with previous section, we insert an asterisk in superscript for $\mathbf{c}_n$, $\mathbf{u}_n$ and $\mathbf{S}_n$ and write

$$\begin{aligned} \mathbf{c}_n^*(\mathbf{x}) &= \mathbf{u}_n^* \, {\rho_n^*}^2(\mathbf{x}) \,, \\ \mathbf{S}_n^* &= \mathbf{u}_n^* {\mathbf{u}_n^*}^\top + 2\, \mathbf{D}_n^{\odot 2}\mathbf{M}^{\odot 2}\mathbf{D}_n^{\odot 2} \,, \end{aligned} \tag{3.17}$$

with

$$\mathbf{u}_n^* = (1/\mathbf{M}_{11}, \ldots, 1/\mathbf{M}_{nn})^\top \,. \tag{3.18}$$

The LOOCV criterion is then equal to

$$\widehat{\mathsf{ISE}}_{LOO}(\eta_n^*) = \frac{1}{n}\sum_{i=1}^n \varepsilon_{-i}^2 = \frac{1}{n}\,\mathbf{y}_n^\top \mathbf{M}\mathbf{D}_n^2\mathbf{M}\mathbf{y}_n \,, \tag{3.19}$$

and its expectation for a given design $\mathbf{X}_n$ is

$$\begin{aligned} \mathsf{E}\left\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n^*)\right\} &= \frac{\sigma^2}{n}\,\mathrm{trace}\left(\mathbf{M}\mathbf{D}_n^2\mathbf{M}\mathbf{K}_n\right) = \frac{\sigma^2}{n}\sum_{i=1}^n \mathbf{M}_{ii}\mathbf{D}_{n\,ii}^2 \\ &= \frac{\sigma^2}{n}\,\mathrm{trace}\left(\mathbf{D}_n\right) = \frac{\sigma^2}{n}\sum_{i=1}^n \frac{1}{\mathbf{M}_{ii}} = \frac{\sigma^2}{n}\sum_{i=1}^n {\rho_{n\setminus i}^*}^2(\mathbf{x}_i) \,. \end{aligned}$$

To shed some light on the issues involved in using (2.5) as a measure of prediction accuracy, let us consider two extreme situations: (i) $\mathbf{X}_n$ is such that each design point has another one in its vicinity; (ii) $\mathbf{X}_n$ is such that all design points are far enough from each other to have a negligible correlation between $Y_{\mathbf{x}_i}$ and $Y_{\mathbf{x}_j}$ for $i \neq j$. In the first case, since when $\mathbf{x}_i$ is dropped there is another design point in its neighborhood, we have $\mathbf{u}_n \simeq 0$, and thus $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ is overoptimistic, with a negative bias close to $-\mathsf{IMSE}(\eta_n)$, see (3.15). The behavior in case (ii) depends on η_n, but for any reasonable predictor such that on average the accuracy of $\eta_n(\mathbf{x})$ improves when $\mathbf{x}$ is closer to a design point $\mathbf{x}_i$ (and thus the correlation between $Y_{\mathbf{x}}$ and $Y_{\mathbf{x}_i}$ increases), on average ε_{-i}^2 will be larger than $\varepsilon_n^2(\mathbf{x})$: $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ will thus tend to be pessimistic, with a positive bias. One may refer to Section 5.1 for an illustrative example.

The actual performance of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ therefore strongly depends on the design configuration: its attractive features pointed out in the literature are in fact valid on average, for designs $\mathbf{X}_n$ with $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mu$. Note that when the $\mathbf{x}_i$ are designed to ensure precise prediction of f over $\mathscr{X}$, they are usually space-filling and thus far from resembling an i.i.d. sample, see, e.g., [17]. This situation approaches case (ii) and we will see in Sections 5.2 to 5.4 that indeed $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ tends to overestimate $\mathsf{ISE}(\eta_n)$.

4 Best linear estimation of the ISE based on squared LOO residuals

The LOOCV estimator (2.5) has two antagonist distinctive features: (i) it is free of any modeling assumption, and can thus be used to infer the predictive quality of any predictor η_n; (ii) it is agnostic to the geometry of the design $\mathbf{X}_n$, and is thus unable to capture its impact on the expected errors in regression problems, being thus highly sensitive to the covariate shift problem [23]. On the one hand, the estimator we propose in this paper loses the universality of feature (i) as it is strongly grounded on the GP assumption for f and is derived for linear predictors only. On the other hand, its parameters are tuned to the design geometry, so that the impact of this geometry on the estimated error is correctly accounted for. A key advantage over LOOCV is thus that we can choose the design $\mathbf{X}_n$ by focusing solely on the accuracy of η_n, without worrying about the possible impact of the choice of $\mathbf{X}_n$ on the precision of the estimate of $\mathsf{ISE}(\eta_n)$.

Our estimator, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, relies on the assumption that $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma_e^2 K^{(e)})$ for some some SPD kernel $K^{(e)}$. In Sections 4.1 to 4.3 we assume that $\mathsf{GP}(0, \sigma_e^2 K^{(e)}) \equiv \mathsf{GP}(0, \sigma^2 K)$ and drop the superscript $^{(e)}$, but in Section 4.4 we investigate the performance (bias and MSE) of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ in the case where the data generating model $\mathsf{GP}(0, \sigma^2 K)$ differs from the assumed model $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$; see also Sections B and C of the supplement.

4.1 Best linear estimation of $\varepsilon_n^2(\mathbf{x})$

We consider estimation of the squared prediction error $\varepsilon_n^2(\mathbf{x})$ of a linear predictor η_n at a generic point $\mathbf{x} \in \mathscr{X}$, based on the n observed squared LOO residuals $\boldsymbol{\varepsilon}_{LOO}^{\odot 2}$, assuming that $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$. The linear estimator that minimizes $\mathsf{E}\{[\varepsilon_n^2(\mathbf{x}) - \boldsymbol{\beta}^\top \boldsymbol{\varepsilon}_{LOO}^{\odot 2}]^2\}$ with respect to $\boldsymbol{\beta} \in \mathbb{R}^n$ is

$$\widehat{\varepsilon}_{nBLP}^2(\mathbf{x}) = \widehat{\boldsymbol{\beta}}^\top(\mathbf{x}) \boldsymbol{\varepsilon}_{LOO}^{\odot 2}, \qquad \text{with} \qquad \widehat{\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{S}_n^{-1} \mathbf{c}_n(\mathbf{x}) \,, \tag{4.1}$$

where $\mathbf{S}_n = \mathsf{E}\{\boldsymbol{\varepsilon}_{LOO}^{\odot 2} {\boldsymbol{\varepsilon}_{LOO}^{\odot 2}}^\top\}/\sigma^4$ and $\mathbf{c}_n(\mathbf{x}) = \mathsf{E}\{\varepsilon_n^2(\mathbf{x}) \boldsymbol{\varepsilon}_{LOO}^{\odot 2}\}/\sigma^4$ are respectively given by (3.5) and (3.6). The assumption that $\mathbf{R}_n$ has full rank implies that $\mathbf{S}_n$ is invertible.

One may notice that when $\mathbf{w}_n(\mathbf{x}_i) = \mathbf{e}_i$, the i-th canonical basis vector, for all $i = 1, \ldots, n$ (which is the case for example when η_n is a kriging predictor for a kernel $K^{(p)}$), then $\rho_n^2(\mathbf{x}_i) = 0$ and $\mathbf{t}_n(\mathbf{x}_i) = \mathbf{0}$ for all i, see (3.2) and (3.3), and therefore $\mathbf{c}_n(\mathbf{x}_i) = \mathbf{0}$, see (3.6), implying that $\widehat{\varepsilon}_{nBLP}^2(\mathbf{x}_i) = 0$ for all i. This is however not necessarily the case for an arbitrary predictor η_n.

The estimate of $\mathsf{ISE}(\eta_n)$ proposed in this paper is

$$\widehat{\mathsf{ISE}}_{BLP}(\eta_n) \;=\; \int_{\mathscr{X}} \widehat{\varepsilon}_{nBLP}^2(\mathbf{x}) \, \mu(\mathrm{d}\mathbf{x}) = {\boldsymbol{\varepsilon}_{LOO}^{\odot 2}}^\top \mathbf{S}_n^{-1} \int_{\mathscr{X}} \mathbf{c}_n(\mathbf{x}) \, \mu(\mathrm{d}\mathbf{x}) = {\boldsymbol{\varepsilon}_{LOO}^{\odot 2}}^\top \mathbf{S}_n^{-1} \mathbf{b}_n \,, \tag{4.2}$$

with $\mathbf{b}_n$ given by (3.12). When μ is approximated by a discrete measure on N points $\mathbf{x}^{(i)}$, $i = 1, \ldots, N$, the complexity of the evaluation of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is of the order $\mathcal{O}(Nn^3)$ ($\mathcal{O}(n^3)$ for the evaluation of each $\widehat{\varepsilon}_{nBLP}^2(\mathbf{x}^{(i)})$). As the minimization of $\mathsf{MSE}\{\widehat{\mathsf{ISE}}(\eta_n)\}$ with respect to $\boldsymbol{\gamma}$ in (3.9) yields $\widehat{\boldsymbol{\gamma}}_{BLP} = \mathbf{S}_n^{-1} \mathbf{b}_n = \int_{\mathscr{X}} \widehat{\boldsymbol{\beta}}(\mathbf{x}) \, \mu(\mathrm{d}\mathbf{x})$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is the best estimator of $\mathsf{ISE}(\eta_n)$ that is linear in the ε_{-i}^2. Note that there is no guarantee that $\widehat{\varepsilon}_{nBLP}^2(\mathbf{x}) = \widehat{\boldsymbol{\beta}}^\top(\mathbf{x}) \boldsymbol{\varepsilon}_{LOO}^{\odot 2}$ be positive. We keep this $\widehat{\varepsilon}_{nBLP}^2(\mathbf{x})$ in our analysis, but use $\widehat{\varepsilon}_{nBLP}^{2\,+}(\mathbf{x}) = \max\{\widehat{\varepsilon}_{nBLP}^2(\mathbf{x}), 0\}$ in the numerical implementation that generated the examples provided in Section 5 and in the

supplement: $\widehat{\varepsilon^2}^{+}_{nBLP}(\mathbf{x})$ minimizes $\mathsf{E}\{[\varepsilon_n^2(\mathbf{x}) - \boldsymbol{\beta}^\top \boldsymbol{\varepsilon}_{LOO}^{\odot 2}]^2\}$ with respect to $\boldsymbol{\beta} \in \mathbb{R}^n$ under the constraint $\boldsymbol{\beta}^\top \boldsymbol{\varepsilon}_{LOO}^{\odot 2} \geq 0$.

4.2 Mean and MSE of the best linear ISE estimator (no modeling error)

The assumption $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$ allows us to compute the statistical moments of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, in particular its bias and MSE. Substituting $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}_{BLP} = \mathbf{S}_n^{-1}\mathbf{b}_n$ in (3.7) and (3.9) we get $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} = \sigma^2 \mathbf{b}_n^\top \mathbf{S}_n^{-1}\mathbf{u}_n$ and thus the bias of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is

$$\mathsf{Bias}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} = \sigma^2 \mathbf{b}_n^\top \mathbf{S}_n^{-1}\mathbf{u}_n - \sigma^2 J_n \,.$$

Its mean squared error is

$$\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} \;=\; \sigma^4 (J_n^2 + 2\,V_n) - \sigma^4 \mathbf{b}_n^\top \mathbf{S}_n^{-1}\mathbf{b}_n \,, \tag{4.3}$$

where $\mathbf{u}_n$, $\mathbf{S}_n$, $\mathbf{b}_n$, J_n and V_n are respectively given by (3.4), (3.5), (3.12), (3.13) and (3.14). Notice that $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} < \sigma^4(J_n^2 + 2\,V_n) = \mathsf{E}\{\mathsf{ISE}^2(\eta_n)\}$, the MSE of the trivial estimator $\widehat{\mathsf{ISE}}(\eta_n) = 0$ — which is not necessarily the case for $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$, see (3.16). Direct comparison with (3.16) gives

$$\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\} - \mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} = \sigma^4 \left(\mathbf{1}_n/n - \mathbf{S}_n^{-1}\mathbf{b}_n\right)^\top \mathbf{S}_n \left(\mathbf{1}_n/n - \mathbf{S}_n^{-1}\mathbf{b}_n\right) \geq 0$$

and Section 5 will highlight the superiority of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ over $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ in various situations involving model misspecification; see also Sections B to D in the supplement.

4.3 BLUP for the assumed GP model

As in Section 3.2, assume now that $\eta_n \equiv \eta_n^*$ given by (2.2) (with again $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$). With the notation of Section 3.2 we get $\mathbf{b}_n^* = \int_{\mathscr{X}} \mathbf{c}_n^*(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) = J_n^* \mathbf{u}_n^*$ with $J_n^* = \int_{\mathscr{X}} {\rho_n^*}^2(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x})$, and our linear estimator has the simple form

$$\widehat{\mathsf{ISE}}_{BLP}(\eta_n^*) = {\boldsymbol{\varepsilon}_{LOO}^{\odot 2}}^\top \widehat{\boldsymbol{\gamma}}_{BLP}^* = J_n^* \, {\boldsymbol{\varepsilon}_{LOO}^{\odot 2}}^\top {\mathbf{S}_n^*}^{-1} \mathbf{u}_n^* \,, \tag{4.4}$$

where $\mathbf{S}_n^*$ and $\mathbf{u}_n^*$ are given by (3.17) and (3.18). Simple algebraic manipulations yield

$$\mathsf{Bias}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n^*)\} \;=\; \sigma^2 J_n^* \left({\mathbf{u}_n^*}^\top {\mathbf{S}_n^*}^{-1} \mathbf{u}_n^* - 1\right) = -\frac{\sigma^2 J_n^*}{1 + {\mathbf{u}_n^*}^\top \mathbf{Q}_n^{-1} \mathbf{u}_n^*} \,,$$

where $\mathbf{Q}_n = 2(\mathbf{D}_n \mathbf{M} \mathbf{D}_n)^{\odot 2}$, showing that $\widehat{\mathsf{ISE}}_{BLP}(\eta_n^*)$ is negatively biased. As the numerical results presented in Section 5 show, this is the case in most situations of interest, and we present a bias-corrected version in Section 4.6. We also get from (4.3):

$$\begin{aligned} \mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n^*)\} &= \mathsf{var}\{\mathsf{ISE}(\eta_n^*)\} + \mathsf{IMSE}^2(\eta_n^*) \left(1 - {\mathbf{u}_n^*}^\top {\mathbf{S}_n^*}^{-1} \mathbf{u}_n^*\right) \\ &= \mathsf{var}\{\mathsf{ISE}(\eta_n^*)\} + \mathsf{IMSE}^2(\eta_n^*) \, \frac{1}{1 + {\mathbf{u}_n^*}^\top \mathbf{Q}_n^{-1} \mathbf{u}_n^*} \,. \end{aligned}$$

4.4 Best linear ISE estimation with model misspecification

In this section, we return to the general framework of Section 4.1, where $\eta_n(\mathbf{x}) = \mathbf{w}^\top(\mathbf{x}, \mathbf{X}_n)\mathbf{y}_n$ is a given arbitrary linear predictor, but we estimate $\mathsf{ISE}(\eta_n)$ assuming the misspecified model $Y_\mathbf{x} \sim \mathsf{GP}(0, \sigma_e^2 K^{(e)})$ when in fact $Y_\mathbf{x} \sim \mathsf{GP}(0, \sigma^2 K)$, $K \neq K^{(e)}$. We thus add the superscript $^{(e)}$ to the notation of Section 4.1. We have now $\widehat{\boldsymbol{\gamma}}_{BLP} = {\mathbf{S}_n^{(e)}}^{-1}\mathbf{b}_n^{(e)}$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n) = {\boldsymbol{\varepsilon}_{LOO}^{\odot 2}}^\top {\mathbf{S}_n^{(e)}}^{-1}\mathbf{b}_n^{(e)}$. Substituting $\widehat{\boldsymbol{\gamma}}_{BLP}$ for $\boldsymbol{\gamma}$ in (3.7) and (3.9) we obtain

$$\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} = \sigma^2\, \mathbf{u}_n^\top {\mathbf{S}_n^{(e)}}^{-1}\mathbf{b}_n^{(e)}\,, \tag{4.5}$$

$$\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} = \sigma^4 \left[{\mathbf{b}_n^{(e)}}^\top {\mathbf{S}_n^{(e)}}^{-1}\mathbf{S}_n {\mathbf{S}_n^{(e)}}^{-1}\mathbf{b}_n^{(e)} - 2\,{\mathbf{b}_n^{(e)}}^\top {\mathbf{S}_n^{(e)}}^{-1}\mathbf{b}_n + J_n^2 + 2\,V_n\right] \tag{4.6}$$

where $\mathbf{u}_n, \mathbf{S}_n, \mathbf{b}_n, J_n$ and V_n are defined in Section 3.1 (with the superscript $^{(e)}$ when the kernel $K^{(e)}$ is substituted for K).

As one may expect, $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ is minimum when using the oracle ISE estimator based on the true model $\mathsf{GP}(0, \sigma^2 K)$. Indeed, denoting $\widehat{\mathsf{ISE}}_{BLP}^{(oracle)}(\eta_n)$ the estimator that uses K instead of $K^{(e)}$, by direct calculation with (4.6) we get

$$\begin{aligned}&\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} - \mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}^{(oracle)}(\eta_n)\} \\ &\qquad = \sigma^4 \left({\mathbf{S}_n^{(e)}}^{-1}\mathbf{b}_n^{(e)} - \mathbf{S}_n^{-1}\mathbf{b}_n\right)^\top \mathbf{S}_n \left({\mathbf{S}_n^{(e)}}^{-1}\mathbf{b}_n^{(e)} - \mathbf{S}_n^{-1}\mathbf{b}_n\right) \geq 0\,.\end{aligned} \tag{4.7}$$

4.5 Independent and flat limits

Here we assume that $K^{(e)}$ is translation invariant, with $K^{(e)}(\mathbf{x}, \mathbf{x}') = K_\theta(\mathbf{x}, \mathbf{x}') = \Psi[\theta(\mathbf{x}-\mathbf{x}')]$, for some function Ψ defined on $\mathbb{R}^+$ such that $\Psi(\mathbf{0}_d) = 1$ and $\Psi(\mathbf{z})$ tending to zero when $\|\mathbf{z}\| \to +\infty$. In particular, K_θ may be isotropic, with $K_\theta(\mathbf{x}, \mathbf{x}') = \psi(\theta\|\mathbf{x}-\mathbf{x}'\|)$ and θ acting like the inverse of a correlation length, with $\psi(0) = 1$ and $\psi(r) \to 0$ as $r \to +\infty$. All kernels used in the examples of Section 5 have this property. We assume that $\mu(\mathbf{X}_n) = 0$. We call independent limit the case $\theta \to +\infty$ and flat limit the case $\theta \to 0$. There is no limiting behaviors to consider for $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ as it does not use $K^{(e)}$, and we thus only consider the case of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ defined by (4.2).

Independent limit For a fixed design $\mathbf{X}_n$, as $\theta \to +\infty$ direct calculation gives $\mathbf{K}_n^{(e)} \to \mathbf{I}_n$, $\mathbf{u}_n^{(e)} \to \mathbf{u}_n^{(e)}(\infty) = \mathrm{diag}\left\{(\mathbf{R}_n^\top \mathbf{R}_n)\right\}$, and we get

$$\begin{aligned}J_n^{(e)} &\underset{\theta\to+\infty}{\to} J_n^{(e)}(\infty) = 1 + \int_\mathscr{X} \|\mathbf{w}_n(\mathbf{x})\|^2\, \mu(\mathrm{d}\mathbf{x})\,,\\ \mathbf{b}_n^{(e)} &\underset{\theta\to+\infty}{\to} \mathbf{b}_n^{(e)}(\infty) = J_n^{(e)}(\infty)\, \mathbf{u}_n^{(e)}(\infty) + 2\,\mathrm{diag}\left\{(\mathbf{R}_n^\top I(\mathbf{w}) \mathbf{R}_n)\right\}\,,\\ \mathbf{S}_n^{(e)} &\underset{\theta\to+\infty}{\to} \mathbf{S}_n^{(e)}(\infty) = \mathbf{u}_n^{(e)}(\infty)\mathbf{u}_n^{(e)}(\infty)^\top + 2\,(\mathbf{R}_n^\top \mathbf{R}_n)^{\odot 2}\,,\end{aligned}$$

with $I(\mathbf{w}) = \int_\mathscr{X} \mathbf{w}_n(\mathbf{x})\mathbf{w}_n^\top(\mathbf{x})\, \mu(\mathrm{d}\mathbf{x})$. Therefore,

$$\widehat{\mathsf{ISE}}_{BLP}(\eta_n) \underset{\theta\to+\infty}{\to} \mathbf{v}_n^\top {\mathbf{S}_n^{(e)}}^{-1}(\infty)\mathbf{b}_n^{(e)}(\infty) \quad \text{and} \quad \mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} \underset{\theta\to+\infty}{\to} \sigma^2\, \mathbf{u}_n^\top \mathbf{S}_n^{(e)}(\infty)^{-1}\mathbf{b}_n^{(e)}(\infty)\,,$$

and, similarly, the independent limit for $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ is obtained by substituting $\mathbf{S}_n^{(e)}(\infty)$ and $\mathbf{b}_n^{(e)}(\infty)$ for $\mathbf{S}_n^{(e)}$ and $\mathbf{b}_n^{(e)}$ in (4.6).

Flat limit We let now θ tend to zero in $K^{(e)}(\mathbf{x},\mathbf{x}') = \Psi[\theta(\mathbf{x}-\mathbf{x}')]$. When we take $K = K^{(e)}$, i.e., when the data are also generated by $\mathsf{GP}(0,\sigma_e^2 K^{(e)})$, a careful analysis (which is beyond the scope of this paper) based on [2, 3] shows the existence of a flat limit for the weights $\widehat{\boldsymbol{\gamma}}^*_{BLP} = J^*_n \, {\mathbf{S}^*_n}^{-1}\mathbf{u}^*_n$ of the estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta^*_n)$. Then, since $\mathsf{E}\{\mathbf{v}_n\} = \sigma^2\mathbf{u}^*_n$ and $\{\mathbf{u}^*_n\}_i = {\rho^*_{n\setminus i}}^2(\mathbf{x}_i)$, see Section 3.2, for a fixed σ_e^2, $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta^*_n)\} \underset{\theta\to 0}{\to} 0$ (and similarly $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta^*_n)\} \underset{\theta\to 0}{\to} 0$).

As explained below, the situation is different in the (more meaningful) situation where $K \neq K^{(e)}$, K is fixed and $\theta \to 0$ in $K^{(e)}$. Studying the precise behavior of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ when $\theta \to 0$ would require developments beyond the scope of this paper; we nevertheless list some basic facts that explain certain features observed in the examples in Section 5.

As $\theta \to 0$, we have $\mathbf{K}^{(e)}_n \to \mathbf{1}_n\mathbf{1}_n^\top$ and $\mathbf{k}^{(e)}_n(\mathbf{x}) \to \mathbf{1}_n$ for all $\mathbf{x}$. Therefore, (3.2) gives ${\rho^{(e)}_n}^2(\mathbf{x}) \underset{\theta\to 0}{\to} [1-\mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2$ and thus $J^{(e)}_n \underset{\theta\to 0}{\to} J^{(e)}_n(0) = \int_{\mathscr{X}}[1-\mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2\,\mu(\mathrm{d}\mathbf{x})$, and (3.3) yields $\mathbf{t}^{(e)}_n(\mathbf{x}) \underset{\theta\to 0}{\to} [1-\mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]\mathbf{1}_n$. We also have

$$\mathbf{u}^{(e)}_n = \mathrm{diag}\left\{(\mathbf{R}_n^\top\mathbf{K}^{(e)}_n\mathbf{R}_n)\right\} \underset{\theta\to 0}{\to} \mathbf{u}^{(e)}_n(0) = (\mathbf{R}_n^\top\mathbf{1}_n)^{\odot 2}\,,$$

so that (3.6) gives $\mathbf{c}^{(e)}_n(\mathbf{x}) \underset{\theta\to 0}{\to} 3\,[1-\mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2\,\mathbf{u}^{(e)}_n(0)$ for any $\mathbf{x}\in\mathscr{X}$, and therefore

$$\mathbf{b}^{(e)}_n \underset{\theta\to 0}{\to} \mathbf{b}^{(e)}_n(0) = 3\,J^{(e)}_n(0)\mathbf{u}^{(e)}_n(0)\,.$$

As $\mathbf{R}_n^\top\mathbf{K}^{(e)}_n\mathbf{R}_n \underset{\theta\to 0}{\to} \mathbf{R}_n^\top\mathbf{1}_n\mathbf{1}_n^\top\mathbf{R}_n$, $\mathbf{S}^{(e)}_n \underset{\theta\to 0}{\to} \mathbf{S}^{(e)}_n(0) = 3\,\mathbf{u}^{(e)}_n(0)[\mathbf{u}^{(e)}_n(0)]^\top$, a rank-one matrix. The singularity of $\mathbf{S}^{(e)}_n(0)$ prevents the existence of a flat-limit for $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ when θ tends to zero, explaining why we encounter numerical difficulties for evaluating $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ for (very) small θ.

When the predictor η_n is such that $\mathbf{w}_n^\top(\cdot)\mathbf{1}_n \equiv 1$ for any n and any $\mathbf{X}_n$, we have $J^{(e)}_n(0) = 0$ and thus $\mathbf{b}^{(e)}_n(0) = \mathbf{0}_n$, and moreover the matrix $\mathbf{R}_n$ in (3.1) satisfies $\mathbf{R}_n^\top\mathbf{1}_n = \mathbf{0}_n$ and thus $\mathbf{u}^{(e)}_n(0) = \mathbf{0}_n$. Therefore, $\mathbf{S}^{(e)}_n$ and $\mathbf{b}^{(e)}_n$ respectively tend to the null matrix and null vector when $\theta_{\mathrm{BLP}} \to 0$, and we may expect the numerical difficulties to be less pronounced for predictors with this property; see the numerical example in Section 5.2.2 for an illustration.

4.6 Best linear unbiased estimation of the ISE

Assuming that $Y_{\mathbf{x}} \sim \mathsf{GP}(0,\sigma^2 K)$, we can easily correct the bias of the linear estimator of $\varepsilon^2_n(\mathbf{x})$ derived in Section 4.1: we minimize $\mathsf{E}\{[\varepsilon^2_n(\mathbf{x}) - \boldsymbol{\beta}^\top\boldsymbol{\varepsilon}^{\odot 2}_{LOO}]^2\}$ with respect to $\boldsymbol{\beta}\in\mathbb{R}^n$ under the constraint $\boldsymbol{\beta}^\top\mathsf{E}\{\boldsymbol{\varepsilon}^{\odot 2}_{LOO}\} = \mathsf{E}\{\varepsilon^2_n(\mathbf{x})\}$. Since $\mathsf{E}\{\varepsilon^2_n(\mathbf{x})\} = \sigma^2\rho^2_n(\mathbf{x})$ and $\mathsf{E}\{\boldsymbol{\varepsilon}^{\odot 2}_{LOO}\} = \sigma^2\mathbf{u}_n$, see (3.2) and (3.4), the constraint is $\boldsymbol{\beta}^\top\mathbf{u}_n = \rho^2_n(\mathbf{x})$, which does not depend on σ^2. The optimal solution to this convex minimization problem is

$$\widehat{\boldsymbol{\beta}}_U(\mathbf{x}) = \mathbf{S}_n^{-1}\left[\mathbf{c}_n(\mathbf{x}) + \frac{\rho^2_n(\mathbf{x}) - \mathbf{u}_n^\top\mathbf{S}_n^{-1}\mathbf{c}_n(\mathbf{x})}{\mathbf{u}_n^\top\mathbf{S}_n^{-1}\mathbf{u}_n}\,\mathbf{u}_n\right]. \tag{4.8}$$

The unbiased version of the linear estimate of $\mathsf{ISE}(\eta_n)$ is then $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n) = \widehat{\boldsymbol{\gamma}}^\top_{BLUP}\boldsymbol{\varepsilon}^{\odot 2}_{LOO}$, with $\widehat{\boldsymbol{\gamma}}_{BLUP} = \int_{\mathscr{X}}\widehat{\boldsymbol{\beta}}_U(\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) = \mathbf{S}_n^{-1}\mathbf{b}_n + (J_n - \mathbf{u}_n^\top\mathbf{S}_n^{-1}\mathbf{b}_n)\mathbf{S}_n^{-1}\mathbf{u}_n/(\mathbf{u}_n^\top\mathbf{S}_n^{-1}\mathbf{u}_n)$ (which we also directly obtain by minimization of $\mathsf{MSE}\{\widehat{\mathsf{ISE}}(\eta_n)\}$ in (3.9) under the constraint $\boldsymbol{\gamma}^\top\mathbf{u}_n = J_n$, see

(3.7) and (3.10)). Its bias, variance and MSE are respectively given by (3.7), (3.8) and (3.9) with $\widehat{\boldsymbol{\gamma}}_{BLUP}$ substituted for $\boldsymbol{\gamma}$. Note that $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ is unbiased only in absence of model misspecification. Cases where $\eta_n \equiv \eta_n^*$ and where a misspecified kernel $K^{(e)} \neq K$ is assumed can be considered similarly to Sections 4.3 and 4.4. A numerical illustration of the performance of $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ is given in Sections 5.2.2 (see Figure 5) and C, D in the supplement, indicating a moderate improvement over $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ (in particular because bias cancellation is only effective in the absence of modeling error).

4.7 Introduction of a nugget effect for noisy observations

Statistical modeling of physical systems usually relies on observations corrupted by noise. Suppose that we observe $y(\mathbf{x}_i) = f(\mathbf{x}_i) + \zeta_i$ at the n design points $\mathbf{x}_i$, where the measurement errors ζ_i are i.i.d. random variables. Assuming that these errors are normal, the developments above for the construction of the ISE estimate $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ remain valid provided we use now a GP model with nugget effect: we assume that $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K_r')$ with K_r' defined by $K_r'(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') + r\,\delta_{\mathbf{x},\mathbf{x}'}$, where $\delta_{\mathbf{x},\mathbf{x}'} = 1$ when $\mathbf{x} = \mathbf{x}'$ and is zero otherwise. The implementation of the method requires knowledge of the nugget effect r. Estimating r from the data $\mathscr{F}_n$ is a possible option (see, e.g., [13, Sect. 4]) that we do not pursue here: it raises several issues, notably of robustness (note that both σ^2 and r need to be estimated), which are worth investigating further. However, the numerical results presented in Section D.2 in the supplement indicate that the performance of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ remains noticeably superior to that of LOOCV even when r is severely misspecified.

5 Numerical experiments

5.1 Influence of the design $\mathbf{X}_n$

This simple example illustrates the discussion at the end of Section 3.1. Here, the function f depends on a single variable $x \in \mathscr{X} = [0, 1]$ and is a realization of a GP: $Y_x \sim \mathsf{GP}(0, \sigma^2 K)$ with $\sigma = 1$ and $K(x, x') = K_{3/2,\theta_0}(\mathbf{x}, \mathbf{x}') = \psi_{3/2,\theta_0}(|x - x'|)$, where $\psi_{3/2,\theta}$ corresponds to the isotropic Matérn 3/2 kernel,

$$\psi_{3/2,\theta}(r) = (1 + \sqrt{3}\,\theta\, r)\exp(-\sqrt{3}\,\theta\, r)\,. \tag{5.1}$$

The predictor η_n is the BLUP for the model $\mathsf{GP}(0, \sigma_p^2 K^{(p)})$, with $K^{(p)}(x, x') = K_{5/2,\theta_p}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_p}(|x - x'|)$ corresponding to the isotropic Matérn 5/2 kernel

$$\psi_{5/2,\theta}(r) \;=\; \Big[1 + \sqrt{5}\,\theta\, r + (5/3)\,\theta^2\, r^2\Big]\exp(-\sqrt{5}\,\theta\, r)\,. \tag{5.2}$$

We take $\theta_0 = 5$ and $\theta_p = 2$ (the predictor thus assumes extra regularity and smoothness). We compare the estimators $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ for a particular realization of Y_x and for a family of n-point designs $\mathbf{X}_n(\delta)$, with $n = 10$, ranging from designs composed of 5 pairs of neighboring points to designs well spread over $\mathscr{X}$: $\mathbf{X}_n(\delta) = \{0, 0.2, 0.4, 0.6, 0.8\} \cup \{\delta, 0.2+\delta, 0.4+\delta, 0.6+\delta, 0.8+\delta\}$, $\delta \in [0.005, 0.1]$ (so that $\delta = \min_{i\neq j} |x_i - x_j|$). The estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ uses the true model $\mathsf{GP}(0, \sigma^2 K)$.

The left panel of Figure 1 shows the realization of Y_x defining $f(x)$ (red solid line) and two predictions $\eta_n(x)$ corresponding to $\mathbf{X}_n(0.015)$ (triangles and dotted line in blue) and $\mathbf{X}_n(0.1)$

(circles and dotted line in green); the design points $\{0, 0.2, 0.4, 0.6, 0.8\}$ (present in $\mathbf{X}_n(0.015)$ and $\mathbf{X}_n(0.1)$) are indicated by red stars. The right panel shows the evolution of the ratios $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)/\mathsf{ISE}(\eta_n)$ (black dotted line with $\circ$) and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)/\mathsf{ISE}(\eta_n)$ (magenta dotted line with $+$), in log scale, as functions of δ, for the particular realization of the left panel. The solid line curves, black with $\triangledown$ and magenta with $\star$, are respectively for $\mathsf{E}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}/\mathsf{IMSE}(\eta_n)$ and $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}/\mathsf{IMSE}(\eta_n)$.

When δ is small, prediction at a removed point x_i is accurate due to the presence of another design point nearby, and the LOO error ε_{-i} is significantly smaller than a typical error $\varepsilon_n(x)$ for $x \in \mathscr{X}$. As a consequence, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ strongly underestimates $\mathsf{ISE}(\eta_n)$. Conversely, for designs corresponding to large δ, removing one x_i leaves a big hole in $\mathbf{X}_n$ and prediction at this x_i is inaccurate: ε_{-i} is thus significantly larger than a typical $\varepsilon_n(x)$ and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ overestimates $\mathsf{ISE}(\eta_n)$. On the opposite, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ gives an acceptable estimation of $\mathsf{ISE}(\eta_n)$ for all values of δ considered.

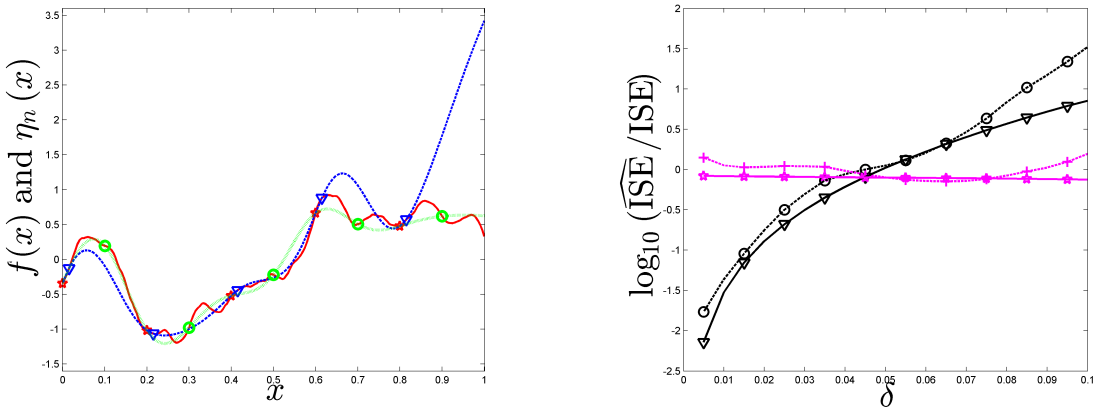

Figure 1: Left: $f(x)$ (—) and $\eta_n(x)$ for the designs $\mathbf{X}_n(0.015)$ ($\cdots$ with $\star$ and $\triangledown$) and $\mathbf{X}_n(0.1)$ ($\cdots$ with $\star$ and $\circ$). Right: $\log_{10}[\widehat{\mathsf{ISE}}_{LOO}(\eta_n)/\mathsf{ISE}(\eta_n)]$ (- - - with $\circ$) and $\log_{10}[\widehat{\mathsf{ISE}}_{BLP}(\eta_n)/\mathsf{ISE}(\eta_n)]$ (- - - with $+$); $\log_{10}[\mathsf{E}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}/\mathsf{IMSE}(\eta_n)]$ (— with $\triangledown$) and $\log_{10}[\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}/\mathsf{IMSE}(\eta_n)]$ (— with $\star$) as functions of δ.

Figure 2 is for the design $\mathbf{X}_n(0.1) = \{0, 0.1, 0.2, \ldots, 0.9\}$ (the best among all $\mathbf{X}_n(\delta)$ in terms of $\mathsf{IMSE}(\eta_n)$) and shows (in log scale) $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\mathsf{ISE}(\eta_n)$ for the realization on the left panel of Figure 1, together with their expected values, when θ_p, the range parameter in the kernel $K_{3/2,\theta_p}$ of the predictor, varies in $[1, 10]$. Estimation of $\mathsf{ISE}(\eta_n)$ by $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is much more accurate than with $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ for all predictors considered. Note that the small negative bias of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is not very sensitive to the smoothness of η_n for this example. Also note that $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ are both minimum for $\theta_p = 10$ whereas $\mathsf{ISE}(\eta_n)$, the true ISE, is minimum for $\theta_p = 1$ (however, $\mathsf{E}\{\mathsf{ISE}(\eta_n)\}$, $\mathsf{E}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$ and $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ are respectively minimum for $\theta_p \simeq 6.6$, 6.7 and 5.9).

5.2 Robustness to the choice of $K^{(e)}$

Here we consider numerical examples of construction of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ involving different predictors η_n, different data generating models $\mathsf{GP}(0, \sigma^2 K)$ and different assumed models $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$

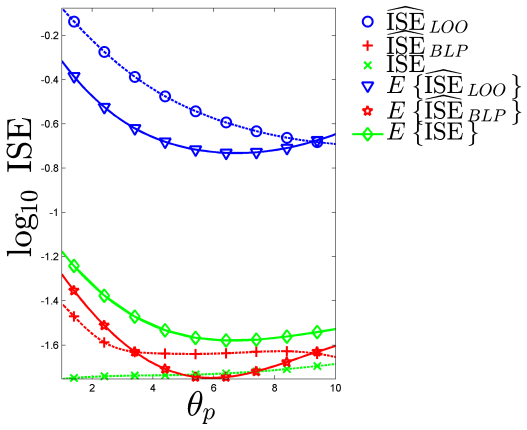

Figure 2: $\log_{10}[\widehat{\mathsf{ISE}}_{LOO}(\eta_n)]$, $\log_{10}[\widehat{\mathsf{ISE}}_{BLP}(\eta_n)]$ and $\log_{10}[\mathsf{ISE}(\eta_n)]$ for the particular realization on the left panel of Figure 1, and $\log_{10}[\mathsf{E}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}]$, $\log_{10}[\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}]$ and $\log_{10}[\mathsf{E}\{\mathsf{ISE}(\eta_n)\}]$, as functions of $\theta_p \in [1, 10]$.

with $K^{(e)} \neq K$, and compare the performances of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$. We always use isotropic kernels. In particular, $K^{(e)}(\mathbf{x}, \mathbf{x}') = \psi^{(e)}(\theta_{\mathrm{BLP}}\|\mathbf{x} - \mathbf{x}'\|)$ and we study the influence of the choice of the range parameter θ_{BLP}, the inverse of a correlation length. The notation θ_{BLP} is to highlight the fact that θ_{BLP} only influences the estimation of $\mathsf{ISE}(\eta_n)$ by $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$.

We take $\mathscr{X} = [0, 1]^2$ and $\mathbf{X}_n$ ($n = 100$) is the 10×10 regular grid with coordinates $(i-1)/9$, $i = 1, \ldots, 10$; μ is the empirical measure on the first 2^{10} points of the low-discrepancy Sobol' sequence in $\mathscr{X}$ (which means that we consider a Quasi-Monte Carlo approximation of the ISE for the uniform measure on $\mathscr{X}$). We generate n observations $\mathbf{y}_n$ for the design $\mathbf{X}_n$ and the model $\mathsf{GP}(0, \sigma^2 K)$, with $\sigma^2 = 1$ (its value is irrelevant as it simple acts as a scaling factor) and $K(\mathbf{x}, \mathbf{x}') = K_{3/2,\theta_0}(\mathbf{x}, \mathbf{x}') = \psi_{3/2,\theta_0}(\|\mathbf{x} - \mathbf{x}'\|)$, the Matérn 3/2 kernel given by (5.1), where we set $\theta_0 = 10$. With θ_0 fixed, the expected ISE for this model, $\mathsf{E}\{\mathsf{ISE}(\eta_n)\} = \mathsf{IMSE}(\eta_n)$, only depends on $\mathbf{X}_n$, see (3.10).

In the first example below, η_n corresponds to a polynomial model that is not an interpolator.

5.2.1 Prediction with a non-interpolating polynomial model

The predictor η_n is obtained by polynomial model fitting: we (wrongly) assume that the data $\mathbf{y}_n$ are given by

$$y_i = \boldsymbol{\phi}^\top(\mathbf{x}_i)\boldsymbol{\alpha} + \delta_i \,,$$

where the error vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ is normally distributed $\mathscr{N}(0, \gamma^2 \mathbf{I}_n)$ and where each component $\phi_\ell(\mathbf{x})$ of $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})]^\top$ is a multivariate polynomial in the d components of $\mathbf{x}$; see Section A in the supplement. The predictor is $\eta_n(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\widehat{\boldsymbol{\alpha}}$, with $\widehat{\boldsymbol{\alpha}}$ the posterior mean of the model parameters and $\boldsymbol{\phi}(\mathbf{x})$ a vector of polynomial functions of $\mathbf{x}$. As this model (wrongly) assumes the presence of i.i.d. observation errors with positive variance γ^2, η_n is not an interpolator.

We take $\gamma^2 = 0.1$; $\boldsymbol{\phi}(\mathbf{x})$ has dimension $m = n/2 = 50$ and each of its components has the form $\varphi_{\ell_1}(x_1)\varphi_{\ell_2}(x_2)$, where $\varphi_i(\cdot)$ denotes the Legendre polynomial of degree i, orthonormal for

the uniform measure on $[0,1]$. The indices ℓ_1 and ℓ_2 take the values

$$\left\{ \begin{bmatrix} \ell_1 \\ \ell_2 \end{bmatrix}, \ell = 1, \ldots, 50 \right\} = \left\{ \begin{array}{l} 00110212032130423140532415062534160735264170845362 \\ 01012021302314032415034251605243617053624718054637 \end{array} \right\},$$

and the polynomial model is thus of total degree 9. We set a vague prior on the parameters $\boldsymbol{\alpha}$, assuming that $\boldsymbol{\alpha} \sim \mathscr{N}(\mathbf{0}_m, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = \mathrm{diag}\{\Lambda_1, \ldots, \Lambda_m\}$ where $\Lambda_\ell = \lambda_{\ell_1}\lambda_{\ell_2}$, $\ell = 1, \ldots, 50$, with ℓ_1, ℓ_2 as indicated above and $\lambda_k = 10^3 \times 2^{-k}$, $k \geq 0$ (the prior on $\boldsymbol{\alpha}$ is therefore not very informative).

We construct $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ using the model $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$, where $K^{(e)} = K_{3/2,\theta_{\mathrm{BLP}}}$ ($K^{(e)}$ thus coincides with K for $\theta_{\mathrm{BLP}} = \theta_0 = 10$). For each value of θ_{BLP} we calculate $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$, respectively given by (4.5) and (4.6).

The procedure is repeated $M = 100$ times, with a different vector $\mathbf{y}_n[i]$ and thus a different predictor $\eta_n[i]$ each time, and we compute the empirical means $\widetilde{\mathsf{E}}\{\mathsf{ISE}(\eta_n)\}$ and $\widetilde{\mathsf{E}}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ of the $\mathsf{ISE}(\eta_n[i])$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n[i])$, respectively, $i = 1, \ldots, M$, together with the empirical standard deviations. We also compute $\widetilde{\mathsf{MSE}}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$, given by the empirical mean of the squared errors $[\mathsf{ISE}(\eta_n[i]) - \widehat{\mathsf{ISE}}_{BLP}(\eta_n[i])]^2$, and use their empirical standard deviation to build confidence intervals.

The left panel of Figure 3 shows $\mathsf{E}\{\mathsf{ISE}(\eta_n)\}$, $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$, $\widetilde{\mathsf{E}}\{\mathsf{ISE}(\eta_n)\}$ and $\widetilde{\mathsf{E}}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ as functions of θ_{BLP}; the confidence bands on $\widetilde{\mathsf{E}}\{\mathsf{ISE}(\eta_n)\}$ and $\widetilde{\mathsf{E}}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ (two standard deviations) are colored. The right panel shows $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ and $\widetilde{\mathsf{MSE}}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ as functions of θ_{BLP}, and the confidence band on $\widetilde{\mathsf{MSE}}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ (two standard deviations, truncated to positive values) is colored. Notice the good agreement between empirical and exact values for the mean and MSE of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$. Although not clearly visible on the plot, $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ is minimum for $\theta_{\mathrm{BLP}} = \theta_0 = 10$, in agreement with (4.7).

On the same data set, the estimator $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ has an empirical mean and standard deviation of approximately 3.6 and 2.5, respectively; its (exact) expected value is about 3.373. These values are well outside the range shown on Figure 3-left. Conversely, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ has a small negative bias for θ_{BLP} around θ_0 or smaller, and its positive bias becomes significant only when θ_{BLP} is much larger than θ_0.

Figure 3 shows that $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ are increasing functions of θ_{BLP} for θ_{BLP} large enough; the independent limits ($\theta_{\mathrm{BLP}} \to +\infty$), obtained from the calculations of Section 4.5, are given in Table 2 and are in good agreement with the values obtained numerically for large θ_{BLP} ($\theta_{\mathrm{BLP}} > 100$, say). These values indicate that $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ gives a much more precise estimation of $\mathsf{ISE}(\eta_n)$ than $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ for all $\theta_{\mathrm{BLP}} \gtrsim 10^{-3}$. Note that $\mathsf{E}\{\mathsf{ISE}^2(\eta_n)\}$ (first column of Table 2) is the MSE of the trivial estimator $\widehat{\mathsf{ISE}}(\eta_n) = 0$ and is much smaller than $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$ (second column).

As the prior on the model parameters is rather vague, the construction almost coincides with Least-Squares regression. Since the linear model contains an intercept ($\phi_1(\mathbf{x}) = 1$ for all $\mathbf{x}$), the prediction $\mathbf{w}_n(\mathbf{x})^\top \mathbf{1}_n$ associated with $\mathbf{y}_n = \mathbf{1}_n$ is almost one for all $\mathbf{x}$: we have $|\mathbf{w}_n(\mathbf{x})^\top \mathbf{1}_n - 1| < 5 \cdot 10^{-8}$ over $\mathscr{X}$. Hence, in agreement with the flat-limit discussion in Section 4.5, small values of θ_{BLP} do not cause severe numerical difficulties, and in Figure 3 we could use values as small as $\theta_{\mathrm{BLP}} = 10^{-3}$. When the prior on $\boldsymbol{\alpha}$ is more informative, the range for θ_{BLP} should be restricted to larger values: for example, when $\lambda_k = 50 \times 2^{-k}$, $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$

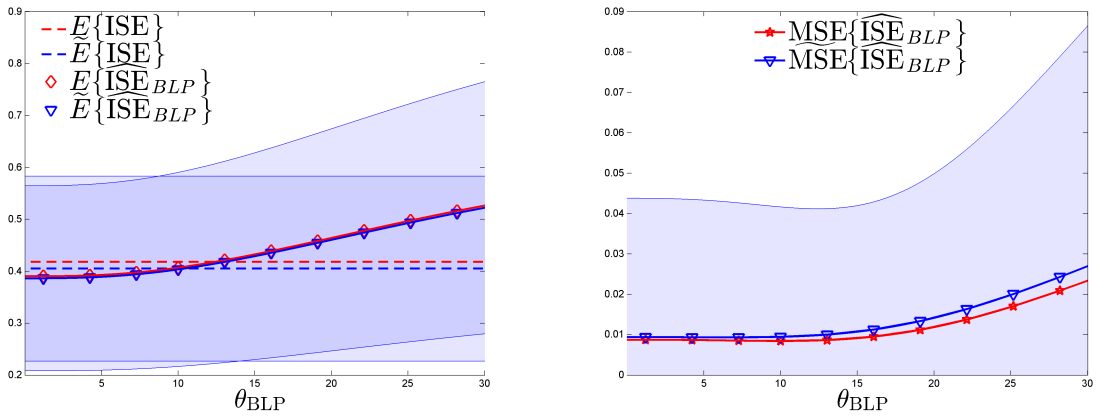

Figure 3: Estimation of $\mathsf{ISE}(\eta_n)$ by $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ when η_n is a (non-interpolating) polynomial of total degree 9 with 100 design points forming a regular grid in $[0,1]^2$; $Y_{\mathbf{x}} \sim \mathsf{GP}(0, K_{3/2,10})$, $K^{(e)} = K_{3/2,\theta_{\mathrm{BLP}}}$, $\theta_{\mathrm{BLP}} \in [0.001, 30]$.

Table 2: $\mathsf{E}\{\mathsf{ISE}(\eta_n)\}$ and $\mathsf{E}\{\mathsf{ISE}^2(\eta_n)\}$ (first column); $\mathsf{E}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$ and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$ (second column); independent limits ($\theta_{\mathrm{BLP}} \to \infty$) for $\mathsf{E}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ (third column). The independent limits are identical for all choices of $K^{(e)}$ considered in the examples of Sections 5.2.2 and B in the supplement.

		$\mathsf{ISE}(\eta_n)$	$\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$	$\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ ($\theta_{\mathrm{BLP}} \to \infty$)
Ex. of Section 5.2.1	E	0.418	3.373	0.672
	MSE	0.181	12.785	0.082
Ex. of Sections 5.2.2 and B	E	0.187	0.731	0.478
	MSE	0.035	0.338	0.103

and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ behave qualitatively like in Figure 3 when $\theta_{\mathrm{BLP}} \gtrsim 0.015$, but numerical instability appears for smaller θ_{BLP}.

5.2.2 Linear prediction with a GP model

The predictor η_n is now the BLUP for the GP model $\mathsf{GP}(0, \sigma_p^2 K^{(p)})$, $\eta_n(\mathbf{x}) = \mathbf{k}_n^{(p)\top}(\mathbf{x}) \mathbf{K}_n^{(p)-1} \mathbf{y}_n$, where $K^{(p)}(\mathbf{x}, \mathbf{x}') = K_{5/2,\theta_p}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_p}(\|\mathbf{x} - \mathbf{x}'\|)$, see (5.2), with $\theta_p = 5$; $\mathbf{X}_n$, $\mathscr{X}$ and μ are like in Section 5.2.1 and the data are still generated with $\mathsf{GP}(0, \sigma^2 K)$ where $\sigma^2 = 1$ and $K = K_{3/2,\theta_0}$ with $\theta_0 = 10$, see (5.1).

We construct $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ for the model $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$, using $K^{(e)} = K_{3/2,\theta_{\mathrm{BLP}}}$ with $\theta_{\mathrm{BLP}} \neq \theta_0$, i.e., $K^{(e)}$ and K have the same regularity but different correlation lengths. Figure 4 presents the same information as Figure 3 in this setting. Comparison of the two figures shows that predictions by the BLUP for the model $\mathsf{GP}(0, \sigma_p^2 K^{(p)})$ are significantly more precise than with the polynomial model of Section 5.2.1. Here, $\widetilde{\mathsf{E}}\{\mathsf{ISE}(\eta_n)\}$ and $\mathsf{E}\{\mathsf{ISE}(\eta_n)\}$ are practically confounded on the left panel; on the right panel, $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ is again minimum for $\theta_{\mathrm{BLP}} = \theta_0 = 10$. In view of the values of $\mathsf{E}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$ and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$ indicated in Table 2, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$

performs significantly better than $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ for the whole range of values of θ_{BLP} considered. Note that the plots are for $\theta_{\mathrm{BLP}} \geq 0.05$ and the numerical difficulties caused by the singularity of the flat limit $\mathbf{S}_n^{(e)}(0)$ of the matrix $\mathbf{S}_n^{(e)}$ are already apparent for θ_{BLP} close to 0.05; see Section 4.5.

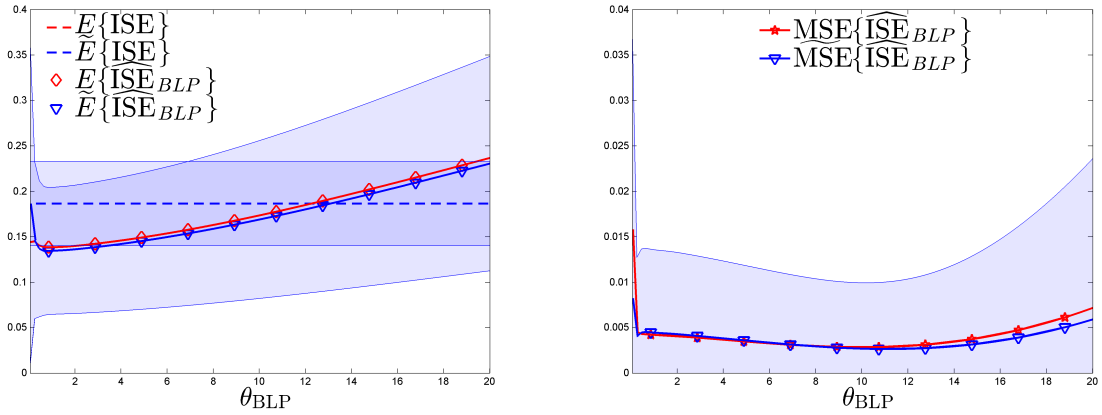

Figure 4: Estimation of $\mathsf{ISE}(\eta_n)$ by $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ when η_n is the BLUP (simple-kriging predictor) for the model $\mathsf{GP}(0, K_{5/2,5})$ on $[0,1]^2$; $Y_{\mathbf{x}} \sim \mathsf{GP}(0, K_{3/2,10})$, $K^{(e)} = K_{3/2,\theta_{\mathrm{BLP}}}$, $\theta_{\mathrm{BLP}} \in [0.05, 20]$ ($\mathbf{X}_n$ is a regular grid of 100 design points).

Figure 5 presents the same information as Figure 4 for the unbiased estimator $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ of Section 4.6. The left panel shows that $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ is indeed unbiased when the model is correct (i.e., for $\theta_{\mathrm{BLP}} = 10$), but remains biased otherwise (and is more biased than $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ for large θ_{BLP}). Its MSE (right panel) is significantly larger (respectively, slightly smaller) than that of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ for large (respectively, small) θ_{BLP}.

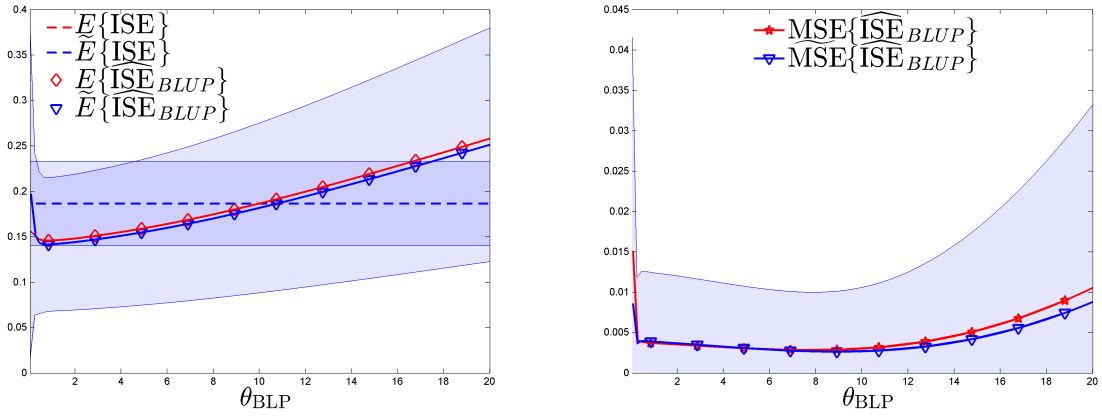

Figure 5: Same as Figure 4 but for the unbiased estimator $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ of Section 4.6.

The behavior of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is slightly different for small θ_{BLP} when η_n is the ordinary-kriging predictor $\widehat{\eta}_n$ for the model $\mathsf{GP}(0, \sigma_p^2 K^{(p)})$. Here, the vector of weights $\widehat{\mathbf{w}}_n(\mathbf{x})$ minimizes

$\rho_n^2(\mathbf{x})$ given by (3.2) for $K = K^{(p)}$ under the constraint $\widehat{\mathbf{w}}_n^\top(\mathbf{x})\mathbf{1}_n = 1$, and is thus solution of

$$\begin{pmatrix} \mathbf{K}_n^{(p)} & \mathbf{1}_n \\ \mathbf{1}_n^\top & 0 \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{w}}_n(\mathbf{x}) \\ \lambda_n \end{pmatrix} = \begin{pmatrix} \mathbf{k}_n^{(p)}(\mathbf{x}) \\ 1 \end{pmatrix},$$

where λ_n is the Lagrange coefficient for the constraint. Denoting by $\overline{\mathbf{K}}_n$ the $(n+1)\times(n+1)$ matrix on the left-hand side and $\overline{\mathbf{M}} = \overline{\mathbf{K}}_n^{-1}$, using block-matrix inversion we get $\varepsilon_{-i} = \overline{\mathbf{M}}_{i,1:n}\mathbf{y}_n/\overline{\mathbf{M}}_{i,i}$ for $i = 1,\ldots,n$, and $\mathbf{R}_n = \overline{\mathbf{M}}\,\overline{\mathbf{D}}_n$ in (3.1), with $\overline{\mathbf{D}}_n = \mathrm{diag}\{1/\overline{\mathbf{M}}_{i,i},\, i = 1,\ldots,n\}$. As $\widehat{\mathbf{w}}_n^\top(\mathbf{x})\mathbf{1}_n = 1$, $\mathbf{R}_n^\top\mathbf{1}_n = \mathbf{0}_n$ and $\mathbf{S}_n^{(e)}$ and $\mathbf{b}_n^{(e)}(\mathbf{x})$ respectively tend to the null matrix and null vector when $\theta_{\mathrm{BLP}} \to 0$. In agreement with the flat-limit discussion in Section 4.5, when θ_{BLP} is small we observe a more stable behavior for $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ on Figure 6 for the ordinary kriging predictor $\widehat{\eta}_n$ than on Figure 4 for the BLUP η_n.

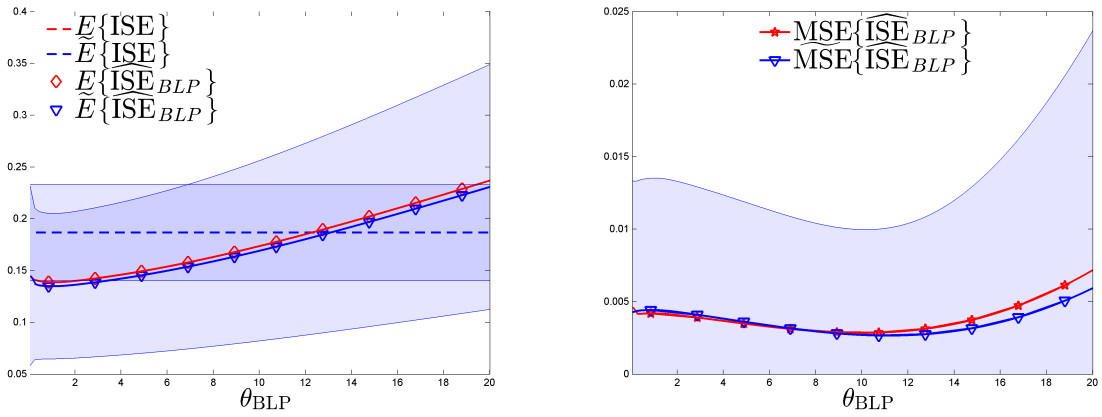

Figure 6: Same as Figure 4 but for the ordinary-kriging predictor $\widehat{\eta}_n$ for the model $\mathsf{GP}(0, K_{5/2,5})$.

In Section B of the supplement, we consider the situation where $K^{(e)}$ and K have different regularities. We still use $K = K_{3/2,10}$ and η_n is the simple-kriging predictor for the model $\mathsf{GP}(0, K_{5/2,5})$, but the construction of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ relies on $K^{(e)}(\mathbf{x},\mathbf{x}') = \psi^{(e)}(\|\mathbf{x}-\mathbf{x}'\|)$, where we consider different $\psi^{(e)}$: $\psi_{1/2,\theta_{\mathrm{BLP}}}(r) = \exp(-\theta_{\mathrm{BLP}}\, r)$, $\psi_{5/2,\theta_{\mathrm{BLP}}}(r)$ given by (5.2), $\psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(r) = (1+\theta_{\mathrm{BLP}}^2\, r^2)^{-1}$ and $\psi_{\infty,\theta_{\mathrm{BLP}}}(r) = \exp(-\theta_{\mathrm{BLP}}^2\, r^2)$, corresponding respectively to the Matérn 1/2, Matérn 5/2, inverse multiquadric and Gaussian kernel. Our conclusion is that the choice of $K^{(e)}$ is not crucial, provided it is regular enough (possibly more regular than K) and θ_{BLP} is not excessively small.

5.3 An environmental model

This example uses the model of [5] that describes the pollutant spill caused by a chemical accident. We use the implementation given at `https://www.sfu.ca/~ssurjano/environ.html`, with parameters set at the values $M = 10$, $D = 0.07$, $L = 1.505$ and $\tau = 30.1525$, as on the figure shown there. The space-time design variables are taken in $\mathscr{X} = [0,3]\times[1,60]$, which we renormalize to $[0,1]^2$. The function varies approximately between 0 and 70 over $\mathscr{X}$, with a rather sharp peak at the center of $\mathscr{X}$.

As the function is fixed, we use random designs to provide a statistical comparison between methods operating in various conditions. We generate random n-point designs in $[0,1]^2$, with $n = 200$, using the relaxed greedy-packing algorithm of [19]. The construction uses $\mathbf{x}_1 = (1/2, 1/2)^\top$ and then $\mathbf{x}_{k+1} = \alpha_k \mathbf{x}_i + (1-\alpha_k)\mathbf{x}^*$ for $k \geq 1$, where $\mathbf{x}^* \in \operatorname{Arg\,max}_{\mathbf{x}\in\mathscr{X}} \min_{\mathbf{x}_j\in\mathbf{X}_k} \|\mathbf{x} - \mathbf{x}_j\|$, $\mathbf{x}_i \in \{\mathbf{x}_\ell \in \mathbf{X}_k : \|\mathbf{x}^* - \mathbf{x}_\ell\| = \min_{\mathbf{x}_j\in\mathbf{X}_k} \|\mathbf{x}^* - \mathbf{x}_j\|\}$, and the α_k are independently uniformly distributed in $[0, a]$, $0 \leq a < 1$. To make the method implementable, we select $\mathbf{x}^*$ within a finite subset $\mathscr{X}_N$ of $\mathscr{X}$: $\mathscr{X}_N$ corresponds to the first $N = 2^{12}$ Sobol' points in $[0,1]^2$. The same points are used to approximate integrals; i.e., μ is the uniform measure on $\mathscr{X}_N$. From [19, Th. 3.6], the packing (respectively, covering) efficiencies of such $\mathbf{X}_n$ with respect to optimal packing (respectively, covering) designs in $\mathscr{X}_N$ equal at least $(1-a)/2$ for any $n \leq N$. We take $a = 0.2$, which yields efficiencies at least 40%.

The predictor for which we estimate the ISE is intentionally not well adapted to this situation: η_n is the BLUP for the model $\mathsf{GP}(0, \sigma_p^2 K^{(p)})$ with $K^{(p)} = K_{3/2,\theta_p}$, see (5.1), i.e., η_n is the simple-kriging predictor for that model. (The ordinary-kriging predictor would be would be a better choice, as the mean of f over $\mathscr{X}_N$ is about 9.5.) The estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ uses the kernel $K^{(e)}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}'\|)$, see (5.2), and different values of θ_{BLP} are considered.

We first set $\theta_p = 1$ in $K^{(p)}$, which provides smooth predictions η_n for designs $\mathbf{X}_n$ well spread over $\mathscr{X}$. The left panel of Figure 7 is for a single (typical) design $\mathbf{X}_n$ and shows $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)/\omega_n$ as a function of θ_{BLP}, with ω_n the empirical variance of the $f(\mathbf{x}_i)$, $\mathbf{x}_i \in \mathbf{X}_n$. Values of $\mathsf{ISE}(\eta_n)/\omega_n$ and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)/\omega_n$ (not depending on θ_{BLP}) are also shown: $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ provides a severe overestimation of $\mathsf{ISE}(\eta_n)$; $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is significantly more accurate for the range considered for θ_{BLP}. The vertical line indicates the value $\widehat{\theta}_{LOO}$, the LOOCV estimator of θ that minimizes $\widehat{\mathsf{ISE}}_{LOO}(\eta_n^*)$, with η_n^* the BLUP for the model $\mathsf{GP}(0, \sigma^2 K_{5/2,\theta})$; see (3.19).

Here, $(1/n)\sum_{i=1}^n y_i \simeq 9.59$ suggesting the use of a GP model with nonzero mean for the construction of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$. However, the weights $\mathbf{w}_n(\mathbf{x})$ of the predictor η_n satisfy $\int_{\mathscr{X}} [1 - \mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2 \, \mu(\mathrm{d}\mathbf{x}) \simeq 2 \cdot 10^{-8}$, and due to the robustness of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ to the presence of a non-zero constant trend when $\mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n \approx 1$ for all $\mathbf{x}$ (see Section E.2 in the supplement), the correction of Section E is not required.

On the right panel of Figure 7, to confirm that the results above are not due to a particularly favorable choice of $\mathbf{X}_n$ we consider 100 random designs (all with packing and covering efficiencies at least 40%). We have seen in Section 5.2 that the estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is not very sensitive to the choice of $K^{(e)}$, suggesting that the precise data fitting of a GP model $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$ is not needed: we use only the isotropic kernel $K^{(e)}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}'\|)$, although other, non-isotropic, kernels may be more suitable; θ_{BLP} is chosen by LOOCV estimation rather than maximum likelihood due the superior robustness of LOOCV to model misspecification, see, e.g., [1], and we take $\theta_{\mathrm{BLP}} = \min\{\max\{\widehat{\theta}_{LOO}, 5\}, 50\}$ (note that $\widehat{\theta}_{LOO}$ is different for each $\mathbf{X}_n$) to compute $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$. The maximum of $\int_{\mathscr{X}} [1 - \mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2 \, \mu(\mathrm{d}\mathbf{x})$ over the 100 designs considered is less than $6 \cdot 10^{-8}$ and we ignore again the presence of a nonzero mean. The figure presents boxplots of the normalized $\mathsf{ISE}(\eta_n)$ (divided by the variance ω_n of the observations $f(\mathbf{x}_i)$, $\mathbf{x}_i \in \mathbf{X}_n$) and of the normalized estimates $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$. Note the much better performance of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ (although significantly worse than that of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$) compared to the example of Section 5.2.1. The computational time[3] for the construction of $\eta_n(\mathbf{x})$ for a

[3]Computations are in Matlab, on a PC with a clock speed of 2.5 GHz and 32 GB RAM.

given $\mathbf{X}_n$ and all $\mathbf{x} \in \mathscr{X}_N$ is about 0.04 s and the calculation of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ takes about 0.09 s (average values over 100 repetitions).

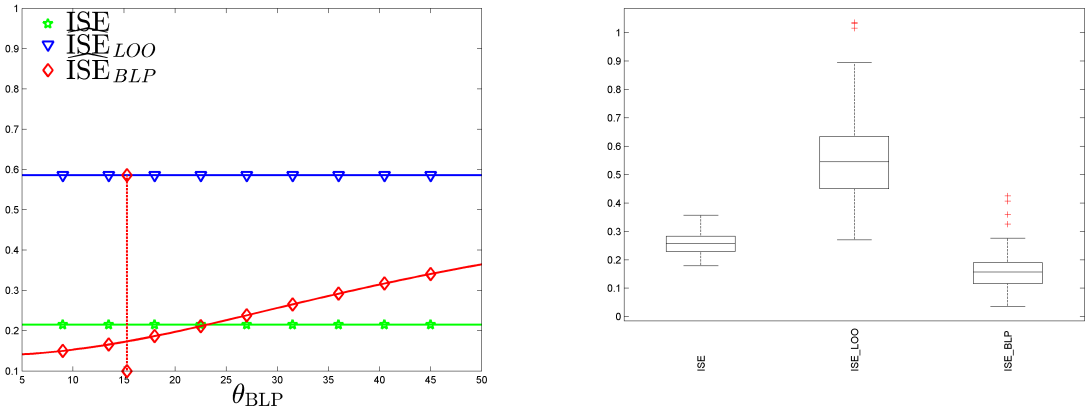

Figure 7: Environmental model: η_n is the BLUP for the model $\mathsf{GP}(0, \sigma^2 K_{3/2,1})$ and $K^{(e)}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}'\|)$. Left: $\mathsf{ISE}(\eta_n)/\omega_n$, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)/\omega_n$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)/\omega_n$ as functions of θ_{BLP}, with ω_n the empirical variance of the $f(\mathbf{x}_i)$, for one random design having packing and covering efficiencies at least 40%; the value $\widehat{\theta}_{LOO}$ that minimizes $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ with η_n the BLUP for the model $\mathsf{GP}(0, \sigma^2 K_{5/2,\theta})$ is indicated by a vertical line. Right: boxplots of $\mathsf{ISE}(\eta_n)/\omega_n$, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)/\omega_n$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)/\omega_n$, for 100 random designs having packing and covering efficiencies at least 40% ($\theta_{\mathrm{BLP}} = \widehat{\theta}_{LOO}$ in $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$).

We use now a predictor with $\theta_p = 1.5546/[2\ \mathsf{PR}(\mathbf{X}_n)]$, with $\mathsf{PR}(\mathbf{X}_n) = (1/2)\ \min_{\mathbf{x}_i \neq \mathbf{x}_j} \|\mathbf{x}_i - \mathbf{x}_j\|$ the packing radius of $\mathbf{X}_n$ (so that $\psi_{3/2,\theta_p}[2\ \mathsf{PR}(\mathbf{X}_n)] \simeq 0.25$, corresponding to a model with rather weak correlation); the shorter correlation length induces a slightly inflated value of $\mathsf{ISE}(\eta_n)$ compared to previous case with $\theta_p = 1$. Figure 8 is the counterpart of Figure 7 for this new situation. On the left panel, the design is the same as in the left panel of Figure 7, we have $\theta_p \simeq 32.8$, $\int_{\mathscr{X}} [1 - \mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2\, \mu(\mathrm{d}\mathbf{x}) \simeq 0.09$ and the performance of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ deteriorates compared with Figure 7 when the nonzero mean is ignored (red curve with diamonds). For all values of θ_{BLP} considered, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ becomes significantly closer to $\mathsf{ISE}(\eta_n)$ when the trend is taken into account via the approach in Section E of the supplement (magenta curve with circles): we assume the model $\mathsf{GP}(\tau, \sigma^2 K_{5/2,\theta_{\mathrm{BLP}}})$, estimate $\mathsf{ISE}_0(\eta_n)$ for centered data as indicated in (E.2), and then add $I(\widehat{\tau}^n)$; see Section E.2. The vertical lines indicate the values of $\widehat{\theta}_{LOO}$ for the two models $\mathsf{GP}(0, \sigma^2 K_{5/2,\theta})$ and $\mathsf{GP}(\tau, \sigma^2 K_{5/2,\theta})$, i.e., with and without zero mean: in the former case, $\widehat{\theta}_{LOO}$ minimizes $\widehat{\mathsf{ISE}}_{LOO}(\eta_n^*)$ with η_n^* the BLUP (the simple kriging predictor) for the model $\mathsf{GP}(0, \sigma^2 K_{5/2,\theta})$; in the second case $\widehat{\theta}_{LOO}$ minimizes $\widehat{\mathsf{ISE}}_{LOO}(\widehat{\eta}_n)$ with $\widehat{\eta}_n$ the ordinary kriging predictor for the model $\mathsf{GP}(\tau, \sigma^2 K_{5/2,\theta})$ — see Section 5.2.2 for the expression of the LOO residuals ε_{-i} in this model.

For the right panel of Figure 8 we use the same 100 random designs as on the right panel of Figure 7 (which gives $\theta_p = 1.5546/[2\ \mathsf{PR}(\mathbf{X}_n)] \in (31, 35)$ for the predictor η_n). As before, we use $\theta_{\mathrm{BLP}} = \min\{\max\{\widehat{\theta}_{LOO}, 5\}, 50\}$ to compute $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, where $\widehat{\theta}_{LOO}$ minimizes either $\widehat{\mathsf{ISE}}_{LOO}(\eta_n^*)$ or $\widehat{\mathsf{ISE}}_{LOO}(\widehat{\eta}_n)$ depending whether we assume a GP with zero mean or not.

Here $0.068 < \int_{\mathscr{X}} [1 - \mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2\, \mu(\mathrm{d}\mathbf{x}) < 0.112$, which is not negligible contrary to previous case with $\theta_p = 1$. When the trend is ignored (third boxplot), $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ performs already much

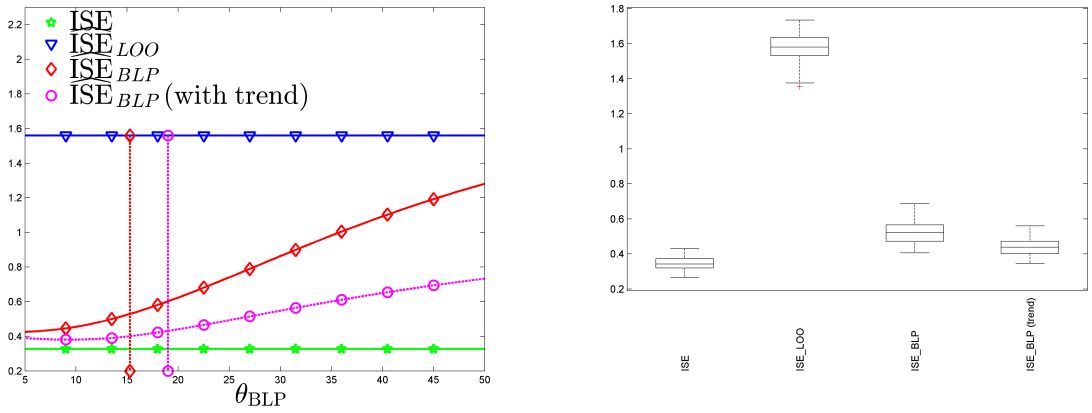

Figure 8: Same as in Figure 7, but η_n is the BLUP for the model $\mathsf{GP}(0, \sigma_p^2 K_{3/2,\theta_p})$ with $\theta_p = 1.5546/[2\ \mathsf{PR}(\mathbf{X}_n)]$. Two versions of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ are considered: one that ignores the nonzero trend, the other that uses the method in Section E of the supplement. On the left panel, the vertical lines indicate the values of $\widehat{\theta}_{LOO}$ for the models $\mathsf{GP}(0, K_{5/2,\theta})$ (◇- - -◇) and $\mathsf{GP}(\tau, K_{5/2,\theta})$ (○- - -○).

better than $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ (second boxplot); performance is further improved when we apply the correction proposed in Section E of the supplement to account for the nonzero mean of $Y_{\mathbf{x}}$, see the fourth boxplot. As the left panel of Figure 8 suggests, better performance could be obtained by choosing a smaller θ_{BLP}. However, optimization with respect to θ_{BLP} is not feasible in a real practical situation as $\mathsf{ISE}(\eta_n)$ is unknown.

We conclude this section by a quick consideration of the problem of model selection. We first highlight that a precise estimator of $\mathsf{ISE}(\eta_n)$ is not an indispensable tool for selecting a predictor from a given class. Indeed, numerical experiments indicate that although $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ is often a poor estimate of $\mathsf{ISE}(\eta_n)$, showing an important positive bias, the predictor that minimizes this estimate has often a small ISE: it is the stability of the precision of the ISE estimate when η_n varies in the class considered that is important, not the absolute precision itself. Hence, although the better performance of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ as an estimator of $\mathsf{ISE}(\eta_n)$ is an invitation to use $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ for model selection, the gain may be marginal.

As an illustration, Figure 9 shows, for the same design $\mathbf{X}_n$ as on the left panels of Figures 7 and 8, the evolution of $\mathsf{ISE}(\eta_n)/\omega_n$, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)/\omega_n$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)/\omega_n$ as functions of θ_p when η_n is the BLUP for the model $\mathsf{GP}(0, \sigma_p^2 K_{3/2,\theta_p})$ (ω_n is the variance of the $f(\mathbf{x}_i)$ for $\mathbf{x}_i \in \mathbf{X}_n$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ uses the correction of Section E of the supplement, with $\theta_{\mathrm{BLP}} = \widehat{\theta}_{LOO}$ for the model $\mathsf{GP}(\tau, \sigma^2 K_{5/2,\theta})$). The estimation of $\mathsf{ISE}(\eta_n)$ by $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is much more precise than by $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ for all values of θ_p considered, but the optimal (minimizing) θ_p for $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ are rather close, with only a slight advantage to the latter (the optimal θ_p being closer to the value minimizing the true ISE, $\mathsf{ISE}(\eta_n)$).

This is confirmed by the results obtained for 100 random designs. We select $\eta_n[\theta_p]$, the BLUP for the model $\mathsf{GP}(0, \sigma^2 K_{3/2,\theta_p})$, among the 46 predictors associated with $\theta_p = 5, 6, \ldots, 50$, by minimization of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n[\theta_p])$ or $\widehat{\mathsf{ISE}}_{BLP}(\eta_n[\theta_p])$. The first row of Table 3 gives, for both estimators, and also for the selection based on the oracle $\mathsf{ISE}(\eta_n[\theta_p])$, the empirical mean of

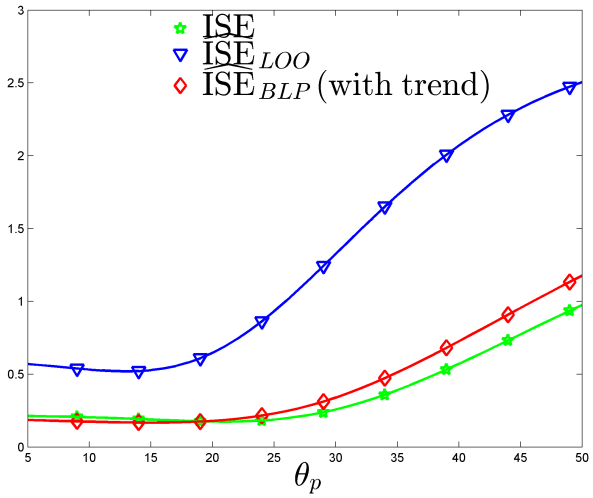

Figure 9: $\mathsf{ISE}(\eta_n)/\omega_n$, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)/\omega_n$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)/\omega_n$, for same design as on the left panels of Figs. 7 and 8, as functions of $\theta_p \in [5, 50]$ (in $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, $\theta_{\text{BLP}} = \widehat{\theta}_{LOO}$ for the model $\mathsf{GP}(\tau, K_{5/2,\theta})$).

$\mathsf{ISE}(\eta_n[\theta_p^{(i)}])/\omega_n$ over the 100 designs, with $\theta_p^{(i)}$ the value associated with the smallest estimated ISE for the i-th design. To appreciate the significance of the numerical values in the table, we also computed the true ISE for the trivial predictor given by $\overline{\eta_n} = \mathbf{y}_n^\top \mathbf{1}_n/n$, i.e., the empirical mean of the observations, and we indicate in the table (last column) the value $(1/100)\sum_{i=1}^{100} \mathsf{ISE}(\overline{\eta_n})/\omega_n$. The estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ uses the trend-correction approach of Section E in the supplement for the model $\mathsf{GP}(\tau, \sigma^2 K_{5/2,\theta_{\text{BLP}}})$ and $\theta_{\text{BLP}} = \widehat{\theta}_{LOO}$ for this model.

We can see that $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ performs slightly better than $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ — but $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ performs surprisingly well if we consider its strong overestimation of the true ISE (it turns out that both estimators very rarely select the same model as the oracle that uses $\mathsf{ISE}(\eta_n[\theta_p])$; see also Figure 1 of Section 5.1 for another illustration).

Table 3: $(1/100)\sum_{i=1}^{100} \mathsf{ISE}(\eta_n[\theta_p^{(i)}])/\omega_n$.

ISE estimator	oracle $\mathsf{ISE}(\eta_n)$	$\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$	$\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$	$\mathsf{ISE}(\overline{\eta_n})$
Ex. of Section 5.3	0.197	0.238	0.224	0.775
Ex. of Section 5.4	$4.13 \cdot 10^{-3}$	$4.47 \cdot 10^{-3}$	$4.23 \cdot 10^{-3}$	0.537

5.4 The piston model

The example concerns a simplified version of a 7-dimensional piston model that describes the motion of a piston within a cylinder, see `https://www.sfu.ca/~ssurjano/piston.html`, with the seven design variables $x_1 = M \in [30, 60]$, $x_2 = S \in [0.005, 0.020]$, $x_3 = V_0 \in [0.002, 0.010]$, $x_4 = k \in [1000, 5000]$, $x_5 = P_0 \in [90000, 110000]$, $x_6 = T_a \in [290, 296]$ and $x_7 = T_0 \in [340, 360]$. As the screening analysis in [15] indicates that only the first four variables have a significant

influence on the model response, we consider a 4-dimensional reduced version of the model, where the input variables $\mathbf{x}_i$ for $i = 5, 6, 7$ are set to the mid-point of the above intervals. The variables $\mathbf{x} = (x_1, \ldots, x_4)$ are renormalized in $\mathscr{X} = [0,1]^4$, we replace $\mathscr{X}$ by the finite set $\mathscr{X}_N$ given by the first $N = 2^{16}$ Sobol' points in $\mathscr{X}$ and take μ equal to the empirical measure on $\mathscr{X}_N$. We then generate random n-point designs in $\mathscr{X}$ (with $n = 50$), using the same greedy-packing algorithm as in Section 5.3, all having packing and covering efficiencies at least 40%. The predictor η_n is again the BLUP for the model $\mathsf{GP}(0, \sigma_p^2\, K_{3/2,\theta_p})$; as the function f is fairly smooth, we take $\theta_p = 1$.

The left panel of Figure 10 presents $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ as functions of θ_{BLP}, together with $\mathsf{ISE}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, for a single design $\mathbf{X}_n$. For the red curve with diamonds, the model assumed is $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$ with $K^{(e)} = K_{5/2,\theta_{\mathrm{BLP}}}$. We have $\int_{\mathscr{X}} [1 - \mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2\, \mu(\mathrm{d}\mathbf{x}) \simeq 0.11 \cdot 10^{-3}$. For large values of θ_{BLP} performance slightly improves when we use the model $\mathsf{GP}(\tau, \sigma^2\, K_{5/2,\theta_{\mathrm{BLP}}})$ with the correction of Section E in the supplement (magenta curve with circles), but the reverse is true for small θ_{BLP}, in particular for $\theta_{\mathrm{BLP}} = \widehat{\theta}_{LOO}$.

This is confirmed by the right panel of Figure 10, which displays boxplots obtained for 100 random designs. We have $0.07 \cdot 10^{-3} < \int_{\mathscr{X}} [1 - \mathbf{w}_n^\top(\mathbf{x})\mathbf{1}_n]^2\, \mu(\mathrm{d}\mathbf{x}) < 0.14 \cdot 10^{-3}$, and the trend-correction of Section E is not quite necessary: on the opposite, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ shows a significantly stronger variability for the model $\mathsf{GP}(\tau, \sigma^2\, K_{5/2,\theta_{\mathrm{BLP}}})$ which accounts for the presence of a nonzero mean than for the model $\mathsf{GP}(0, \sigma^2\, K_{5/2,\theta_{\mathrm{BLP}}})$ ($\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ uses $\theta_{\mathrm{BLP}} = \widehat{\theta}_{LOO}$ for the model considered). Nevertheless, both estimators perform much better than $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$. Note the negligible variability of $\mathsf{ISE}(\eta_n)$ across designs due to the strong regularity of f.

For a given $\mathbf{X}_n$, the computational time of the construction of $\eta_n(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}_N$ is about 0.18 s and the calculation of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ takes about 0.35 s (average values over 100 repetitions).

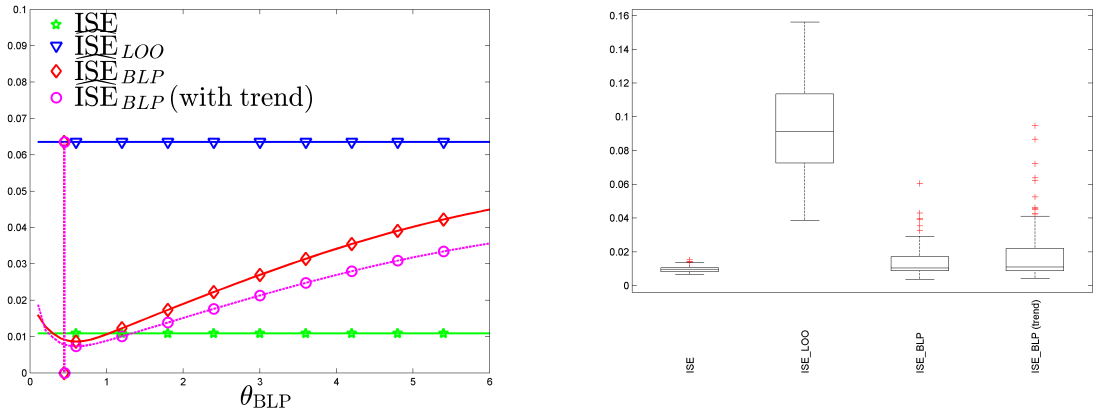

Figure 10: Same as in Figure 8 for the piston model, with η_n the BLUP for $\mathsf{GP}(0, \sigma^2\, K_{3/2,1})$. On the left panel, the values of $\widehat{\theta}_{LOO}$ for the models $\mathsf{GP}(0, K_{5/2,\theta})$ (◇- - -◇) and $\mathsf{GP}(\tau, K_{5/2,\theta})$ (○- - -○), indicated by vertical lines, are practically confounded.

The performance of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ (based on the model $\mathsf{GP}(0, K_{5/2,\widehat{\theta}_{LOO}})$) for model selection is summarized in the second row of Table 3. As in Section 5.3, we select $\eta_n[\theta_p]$ in

a finite family: $\eta_n[\theta_p]$ is the BLUP for the model $\mathsf{GP}(0, \sigma^2\, K_{3/2,\theta_p})$ with $\theta_p \in \{0.01, 0.02, \ldots, 0.5\}$ (50 elements). As for the example of Section 5.3, the predictors selected with $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ have, on average, a slightly smaller ISE than those selected with $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$.

6 Conclusions and further developments

The paper proposes a method that set weights on LOO squared residuals when estimating the ISE of a linear predictor η_n. The resulting ISE estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is more precise than usual (unweighted) LOOCV, sometimes considerably so. The dependence of the weights on the sampling design gives the estimator a certain robustness to the design configuration, unlike LOOCV. On the downside, the method is not as universal as LOOCV: it is limited to ISE estimation for linear predictors and relies on a GP model (or a mixture of GP models) for the function that the predictor approximates. The numerical examples presented indicate reasonable robustness with respect to the choice of the kernel $K^{(e)}$ of the assumed GP model.

Here we have only considered LOO residuals, but the results in [11] open the way to extension to multiple-fold CV. There, the i-th LOO residual ε_{-i} is replaced by a vector of residuals $\boldsymbol{\varepsilon}_{I_i}$ at the design points $\mathbf{X}_{I_i} = \{\mathbf{x}_j,\, j \in I_i\}$ with $I_i \subset \{1, \ldots, n\}$, for which only the other points in $\mathbf{X}_n \setminus \mathbf{X}_{I_i}$ are used for prediction. The weighted ISE estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ for such multiple-fold CV would rely on the construction of the best linear estimator of $\varepsilon_n^2(\mathbf{x})$ based on squared residuals $\boldsymbol{\varepsilon}_{I_i}^{\odot 2}$ for all I_i considered. Under a GP model assumption, the $\boldsymbol{\varepsilon}_{I_i}$ are Gaussian, and the expressions given in [11] can be used to calculate the expectations needed to compute $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, following the same lines as in Section 4.1. When the sets I_i form a partition of $\{1, \ldots, n\}$, the concatenation of the $\boldsymbol{\varepsilon}_{I_i}$ forms a vector of length n, which entails not major changes compared with the developments in Section 4; however, when the concatenation forms a vector of length $m > n$, the associated matrix $\mathbf{S}_m$ is singular and some adaptation becomes necessary.

One may note that when $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$, $\mathsf{E}\{\mathsf{ISE}(\eta_n)\} = \sigma^2 J_n$, see (3.10). This observation prompts us to estimate $\mathsf{ISE}(\eta_n)$ by $\widehat{\sigma}^2 J_n$, with $\widehat{\sigma}^2$ an estimator of the process variance σ^2. In particular, assuming that the data are generated with the model $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$, we may use the maximum-likelihood estimator $\widehat{\sigma}^2_{ML} = (1/n)\mathbf{y}_n^\top {\mathbf{K}_n^{(e)}}^{-1} \mathbf{y}_n$, or the LOO estimator,

$$\widehat{\sigma}^2_{LOO} = \frac{1}{n} \sum_{i=1}^n \mathbf{M}^{(e)}_{ii} \varepsilon^2_{-i} = \frac{1}{n}\, \mathbf{y}_n^\top \left(\sum_{i=1}^n \frac{\mathbf{M}^{(e)}_{\cdot i} \mathbf{M}^{(e)}_{i\cdot}}{\mathbf{M}^{(e)}_{ii}} \right) \mathbf{y}_n = \frac{1}{n}\, \mathbf{y}_n^\top \mathbf{M}^{(e)} \mathbf{D}^{(e)}_n \mathbf{M}^{(e)} \mathbf{y}_n \,,$$

where $\mathbf{M}$ and $\mathbf{D}_n$ are defined in Section 3.2; see, e.g., [6, 1, 14]. Moreover, since $\mathsf{E}\{\mathsf{ISE}(\eta_n^*)\} = \sigma^2 J_n^*$ when $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K)$, exploitation of the expression (4.4) for $\widehat{\mathsf{ISE}}_{BLP}(\eta_n^*)$ suggests that we could also estimate σ^2 by the best linear estimator based on squared LOO residuals, $\widehat{\sigma}^2_{BLP} = \widehat{\mathsf{ISE}}_{BLP}(\eta_n^*)/{J_n^{(e)}}^*$ (or the unbiased version $\widehat{\sigma}^2_{BLUP} = \widehat{\mathsf{ISE}}_{BLUP}(\eta_n^*)/{J_n^{(e)}}^*$). For lack of space, we have not reported here the numerical results obtained with the ISE estimators $\widehat{\sigma}^2_{ML} J_n$, $\widehat{\sigma}^2_{LOO} J_n$ and $\widehat{\sigma}^2_{BLP} J_n$, nor have we presented a comparative study of the performances of $\widehat{\sigma}^2_{ML}$, $\widehat{\sigma}^2_{LOO}$ and $\widehat{\sigma}^2_{BLP}$ as estimators of σ^2 (one may refer to [1] for a comparison between $\widehat{\sigma}^2_{ML}$ and $\widehat{\sigma}^2_{LOO}$) and we content ourselves with delivering the raw conclusion of our observations: $\widehat{\sigma}^2_{BLP}$ is often a valid alternative to $\widehat{\sigma}^2_{ML}$ and $\widehat{\sigma}^2_{LOO}$ (in the same way as $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is a valid alternative to $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$), but the associated ISE estimators are generally not competitive compared with $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, even if they sometimes perform significantly better than $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$.

Finally, we have presented some preliminary, but promising, results concerning the application of our ISE estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ to model selection, in particular to the selection of a GP model when η_n is the BLUP for $\mathsf{GP}(0, \sigma_p^2 K^{(p)})$. In that case, if we take $K^{(p)}$ and $K^{(e)}$ in the same family, the method can be iterated, following the same fixed-point principle as for iteratively reweighted least-squares (see, e.g., [12]): the first kernel K_1 can be initialized through selection by LOOCV; then, at each iteration $j \geq 1$, model selection by minimization of the estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n[K^{(p)}])$ constructed with $K^{(e)} = K_j$, yields the kernel K_{j+1} to be used to calculate $\widehat{\mathsf{ISE}}_{BLP}(\eta_n[K^{(p)}])$ at next iteration. We do not expand on this iterative approach here, although it would be worth exploring further.

The limitations of the method are those inherent in the use of GP models for function approximation. In situations where the predictor η_n under consideration performs well enough, finding an appropriate GP model for f seems to be a feasible task, making $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ a useful tool for estimating $\mathsf{ISE}(\eta_n)$. However, there are situations where η_n performs poorly and where it is difficult to find a suitable GP model for f; in particular, the design $\mathbf{X}_n$ may be too sparse to detect the variability of f (which can happen especially when d is large). If η_n is very inaccurate but has sufficient variability, it is possible that the LOO residuals are large enough for the inaccuracy to be detected by the high value of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ (and possibly of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$). However, if η_n is much smoother than f, it may produce very small LOO residuals and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ may then severely underestimate $\mathsf{ISE}(\eta_n)$ — and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ will not do any better, the principle of LOOCV itself being ineffective. A simple illustrative example (with $d = 1$) is presented in Section G of the supplement. In this case, only the use of an independent test set can help reveal the poor performance of η_n (and, as proposed in [8, 18], ISE estimation can rely on the BLP of the squared errors $\varepsilon_n^2(\mathbf{x})$ based on the squared test residuals). The Matlab code of a function that calculates $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ for a linear predictor η_n is given in Section H of the supplement.

References

[1] F. Bachoc, *Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification*, Comput. Statist. Data Anal., 66 (2013), pp. 55–69.

[2] S. Barthelmé, P.-O. Amblard, N. Tremblay, and K. Usevich, *Gaussian process regression in the flat limit*, The Annals of Statistics, 51 (2023), pp. 2471–2505.

[3] S. Barthelmé and K. Usevich, *Spectral properties of kernel matrices in the flat limit*, SIAM Journal on Matrix Analysis and Applications, 42 (2021), pp. 17–57.

[4] S. Bates, T. Hastie, and R. Tibshirani, *Cross-validation: what does it estimate and how well does it do it?*, Journal of the American Statistical Association, (2023), pp. 1–12.

[5] N. Bliznyuk, D. Ruppert, C. Shoemaker, R. Regis, S. Wild, and P. Mugunthan, *Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation*, Journal of Computational and Graphical Statistics, 17 (2008), pp. 270–294.

[6] N. Cressie, *Statistics for Spatial Data*, Wiley, New York, 1993.

[7] O. Dubrule, *Cross validation of kriging in a unique neighborhood*, Journal of the International Association for Mathematical Geology, 15 (1983), pp. 687–699.

[8] E. Fekhari, B. Iooss, J. Muré, L. Pronzato, and M.-J. Rendas, *Model predictivity assessment: incremental test-set selection and accuracy evaluation*, in Studies in Theoretical and Applied Statistics, N. Salvati, C. Perna, S. Marchetti, and R. Chambers, eds., Springer (Proceedings in Mathematics & Statistics 406), 2022.

[9] B. Gauthier and L. Pronzato, *Spectral approximation of the IMSE criterion for optimal designs in kernel-based interpolation models*, SIAM/ASA J. Uncertainty Quantification, 2 (2014), pp. 805–825.

[10] B. Gauthier and L. Pronzato, *Approximation of IMSE-optimal designs via quadrature rules and spectral decomposition*, Communications in Statistics – Simulation and Computation, 45 (2016), pp. 1600–1612.

[11] D. Ginsbourger and C. Schärer, *Fast calculation of Gaussian process multiple-fold cross-validation residuals and their covariances*, arXiv preprint arXiv:2101.03108v2, (2022).

[12] P. Green, *Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion)*, Journal of the Royal Statistical Society, B-46 (1984), pp. 149–192.

[13] T. Karvonen and C. Oates, *Maximum likelihood estimation in Gaussian process regression is ill-posed*, Journal of Machine Learning Research, 24 (2023), pp. 1–47.

[14] T. Karvonen, G. Wynne, F. Tronarp, C. Oates, and S. Särkkä, *Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions*, SIAM/ASA Journal on Uncertainty Quantification, 8 (2020), pp. 926–958.

[15] H. Moon, *Design and analysis of computer experiments for screening input variables*, PhD thesis, The Ohio State University, 2010.

[16] M. Naslidnyk, M. Kanagawa, T. Karvonen, and M. Mahsereci, *Comparing scale parameter estimators for Gaussian process regression: cross validation and maximum likelihood*, arXiv preprint arXiv:2307.07466, (2023).

[17] L. Pronzato and W. Müller, *Design of computer experiments: space filling and beyond*, Statistics and Computing, 22 (2012), pp. 681–701.

[18] L. Pronzato and M.-J. Rendas, *Validation of machine learning prediction models*, New England J. of Statistics in Data Science, 1 (2023), pp. 394–414.

[19] L. Pronzato and A. Zhigljavsky, *Quasi-uniform designs with asymptotically optimal and near-optimal uniformity constant*, Journal of Approximation Theory, 294 (2023).

[20] J. Sacks, S. Schiller, and W. Welch, *Designs for computer experiments*, Technometrics, 31 (1989), pp. 41–47.

[21] M. Stein, *Interpolation of Spatial Data. Some Theory for Kriging*, Springer, Heidelberg, 1999.

[22] M. Stone, *Cross-validatory choice and assessment of statistical predictions*, Journal of the Royal Statistical Society, Series B (Methodological), 36 (1974), pp. 111–147.

[23] M. Sugiyama, M. Krauledat, and K.-R. Müller, *Covariate shift adaptation by importance weighted cross validation.*, Journal of Machine Learning Research, 8 (2007).

[24] H. Xu and R. Tibshirani, *Estimation of prediction error with known covariate shift*, arXiv preprint arXiv:2205.01849, (2022).

Appendix: supplementary material

A A polynomial regression model

The model assumes that the data $\mathbf{y}_n$ are given

$$y_i = \boldsymbol{\phi}^\top(\mathbf{x}_i)\boldsymbol{\alpha} + \delta_i\,,$$

where the error vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ is normally distributed $\mathscr{N}(0, \boldsymbol{\Omega}_n)$ and where each component $\phi_\ell(\mathbf{x})$ of $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})]^\top$ is a multivariate polynomial in the d components of $\mathbf{x}$. We set a normal prior on $\boldsymbol{\alpha}$, and assume that $\boldsymbol{\alpha} \sim \mathscr{N}(\mathbf{0}_m, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = \mathrm{diag}\{\Lambda_1, \ldots, \Lambda_m\}$. The posterior mean of $\boldsymbol{\alpha}$ under these assumptions is $\widehat{\boldsymbol{\alpha}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Omega}_n^{-1}\boldsymbol{\Phi} + \boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Omega}_n^{-1}\mathbf{y}_n$, where $\boldsymbol{\Phi}$ is the $n \times m$ matrix with i-th row equal to $\boldsymbol{\phi}^\top(\mathbf{x}_i)$, $i = 1, \ldots, n$. The prediction at any given $\mathbf{x}$ is then

$$\eta_n(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\widehat{\boldsymbol{\alpha}} = \boldsymbol{\phi}^\top(\mathbf{x})(\boldsymbol{\Phi}^\top\boldsymbol{\Omega}_n^{-1}\boldsymbol{\Phi} + \boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Omega}_n^{-1}\mathbf{y}_n\,.$$

Straightforward matrix manipulation shows that

$$\eta_n(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\Lambda}\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^\top + \boldsymbol{\Omega}_n)^{-1}\mathbf{y}_n\,.$$

Take $\boldsymbol{\Omega}_n = \gamma^2\,\mathbf{I}_n$, with $\mathbf{I}_n$ the n-dimensional identity matrix. Then, denoting $\mathbf{K}_n^{(p)} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^\top + \gamma^2\,\mathbf{I}_n$ and $\mathbf{k}_n^{(p)}(\mathbf{x}) = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\phi}(\mathbf{x})$, we get $\eta_n(\mathbf{x}) = [\mathbf{k}_n^{(p)}(\mathbf{x})]^\top\mathbf{K}_n^{(p)^{-1}}\mathbf{y}_n$; that is, η_n is the BLUP for a GP model with a kernel with nugget effect, defined by $K^{(p)}(\mathbf{x}, \mathbf{x}') = K^{\phi}(\mathbf{x}, \mathbf{x}') + \gamma^2\delta_{\mathbf{x},\mathbf{x}'}$, where $\delta_{\mathbf{x},\mathbf{x}'} = 1$ when $\mathbf{x} = \mathbf{x}'$ and is zero otherwise and

$$K^{\phi}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=1}^{m} \Lambda_\ell\phi_\ell(\mathbf{x})\phi_\ell(\mathbf{x}')\,. \tag{A.1}$$

Unless $\gamma^2 = 0$ (and $m \geq n$), the predictor η_n is not an interpolator. The position of i in $\{1, \ldots, n\}$ is irrelevant to compute the LOO error $\varepsilon_{-i} = y_i - \eta_{n\setminus i}(\mathbf{x}_i)$, and one may thus consider the case $i = n$. The i-th row of $\mathbf{K}_n^{(p)^{-1}} = (\mathbf{K}_n^{\phi} + \gamma^2\mathbf{I}_n)^{-1}$ then equals

$$\left\{(\mathbf{K}_n^{\phi} + \gamma^2\mathbf{I}_n)^{-1}\right\}_{i\cdot} = \begin{pmatrix} -\frac{[\mathbf{k}_{n\setminus i}^{(p)}(\mathbf{x}_i)]^\top(\mathbf{K}_{n\setminus i}^{\phi} + \gamma^2\mathbf{I}_{n-1})^{-1}}{A_i} & \frac{1}{A_i} \end{pmatrix}$$

with $A_i = K^{(p)}(\mathbf{x}_i,\mathbf{x}_i) - [\mathbf{k}^{(p)}_{n\setminus i}(\mathbf{x}_i)]^\top (\mathbf{K}^{\phi}_{n\setminus i} + \gamma^2 \mathbf{I}_{n-1})^{-1} \mathbf{k}^{(p)}_{n\setminus i}(\mathbf{x}_i)$. We can then identify the i-th row $\mathbf{r}_i^\top$ of the matrix $\mathbf{R}_n$ for the construction of ε_{-i} in (3.1):

$$\begin{aligned}\varepsilon_{-i} = y_i - \eta_{n\setminus i}(\mathbf{x}_i) &= y_i - [\mathbf{k}^{(p)}_{n\setminus i}(\mathbf{x}_i)]^\top (\mathbf{K}^{\phi}_{n\setminus i} + \gamma^2 \mathbf{I}_{n-1})^{-1} \mathbf{y}_{n\setminus i} \\ &= \frac{\left\{(\mathbf{K}^{(p)}_n)^{-1}\right\}_{i\cdot}\mathbf{y}_n}{\left\{(\mathbf{K}^{(p)}_n)^{-1}\right\}_{ii}} = \mathbf{r}_i^\top \mathbf{y}_n\,.\end{aligned}$$

In the example of Section 5.2.1, the polynomial model is constructed by tensorization of univariate polynomials. The index ℓ of a component $\phi_\ell(\mathbf{x})$ of $\boldsymbol{\phi}(\mathbf{x})$ is in fact a multiindex $\underline{\ell} = \{\ell_1,\ldots,\ell_d\}$, with $\phi_{\underline{\ell}}(\mathbf{x}) = \prod_{i=1}^d \varphi_{\ell_i}(x_i)$ for $\mathbf{x} = (x_1,\ldots,x_d)^\top$. The degree of the φ_k increases with k; a scalar λ_k is attached to each of them, and $\Lambda_{\underline{\ell}} = \prod_{i=1}^d \lambda_{\ell_i}$ with λ_k decreasing with k in order to give more importance to lower degree polynomials. Only the terms corresponding to the m largest $\Lambda_{\underline{\ell}}$ is kept to form the kernel (A.1). The construction used below relies on Legendre polynomials, orthonormal for the uniform measure on $[0,1]$: $\varphi_0(x) = 1$, $\varphi_1(x) = \sqrt{3}(2x-1)$, $\varphi_2 = \sqrt{5}(6x^2-6x+1)$, $\varphi_3(x) = \sqrt{7}(20x^3-30x^2+12x-1)\ldots$ As φ_k has degree k, setting $\lambda_k = t^{-k}$ for some $t>1$ gives $\Lambda_{\underline{\ell}} = t^{-\sum_{i=1}^d \ell_i}$ and thus implies that the terms selected in (A.1) are among those with lowest total degree. One may refer to [2] for implementation details.

B Robustness of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ to the choice of $K^{(e)}$: $K^{(e)}$ and K have different regularities

This is a continuation of Section 5.2.2. We still use $K = K_{3/2,10}$ and η_n is the simple-kriging predictor for the model $\mathsf{GP}(0, K_{5/2,5})$, but the construction of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ relies on $K^{(e)}(\mathbf{x},\mathbf{x}') = \psi^{(e)}(\|\mathbf{x}-\mathbf{x}'\|)$, where we consider different $\psi^{(e)}$:

$$\begin{aligned}\psi_{1/2,\theta_{\mathrm{BLP}}}(r) &= \exp(-\theta_{\mathrm{BLP}}\, r)\,, \\ \psi_{5/2,\theta_{\mathrm{BLP}}}(r) &= \left[1+\sqrt{5}\,\theta_{\mathrm{BLP}}\, r + (5/3)\,\theta^2_{\mathrm{BLP}}\, r^2\right]\exp(-\sqrt{5}\,\theta_{\mathrm{BLP}}\, r)\,, \qquad \text{(B.1)} \\ \psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(r) &= (1+\theta^2_{\mathrm{BLP}}\, r^2)^{-1}\,, \qquad \text{(B.2)} \\ \psi_{\infty,\theta_{\mathrm{BLP}}}(r) &= \exp(-\theta^2_{\mathrm{BLP}}\, r^2)\,,\end{aligned}$$

corresponding respectively to the Matérn 1/2, Matérn 5/2, inverse multiquadric and Gaussian kernel.

Figure 11 shows how $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ behaves when θ_{BLP} varies in the four kernels $K^{(e)}$ considered (the behavior for the Matérn 3/2 kernel $K_{3/2,\theta_{\mathrm{BLP}}}$ has already been illustrated in Figure 4). Unsurprisingly, the more regular $K^{(e)}$ is, the stronger is the numerical instability for small θ_{BLP}. The independent limits (for $\theta_{\mathrm{BLP}} \to +\infty$) are nevertheless practically identical for the four choices of $K^{(e)}$ (see Table 2). The choice of $K^{(e)}$ does not appear to be essential, provided it is regular enough (possibly more regular than K) and θ_{BLP} is not excessively small. For each kernel considered, there is a reasonably large range of values of θ_{BLP} such that $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ provides an accurate estimate of $\mathsf{ISE}(\eta_n)$ (compare with the values of $\mathsf{E}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$ and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{LOO}(\eta_n)\}$ given in Table 2), the most stable performance being achieved for $K^{(e)} = K_{5/2,\theta_{\mathrm{BLP}}}$.

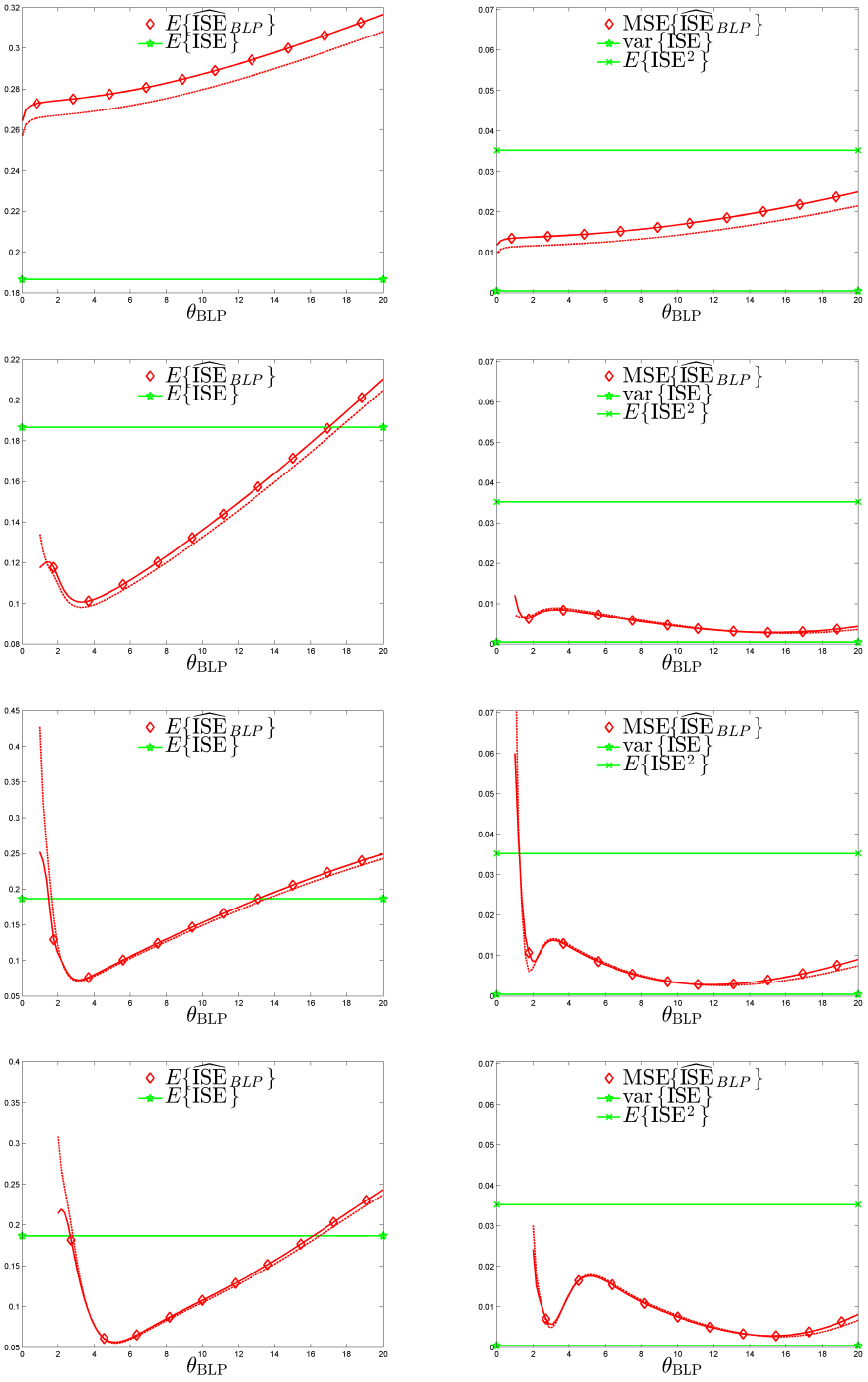

Figure 11: Performance of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ when η_n is the simple-kriging predictor for the model $\mathsf{GP}(0, K_{5/2,5})$ on $[0,1]^2$; $Y_{\mathbf{x}} \sim \mathsf{GP}(0, K_{3/2,10})$, ($\mathbf{X}_n$ is a regular grid of 100 design points). From top to bottom: $\psi^{(e)} = \psi_{1/2,\theta_{\mathrm{BLP}}}$, $\psi_{5/2,\theta_{\mathrm{BLP}}}$, $\psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}$ and $\psi_{\infty,\theta_{\mathrm{BLP}}}$. Empirical values for 100 repetitions are in dotted lines.

C Average performance of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ for GP realizations with $d \in \{4, 6, 8\}$

In this section we present the values of $\mathsf{E}\{\mathsf{ISE}(\eta_n)\}$, $\mathsf{E}\{\widehat{\mathsf{ISE}}(\eta_n)\}$ and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}(\eta_n)\}$ for three different ISE estimators, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ (2.5), $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ (4.2) and its unbiased version $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ of Section 4.6, when f is the realization of a GP. The design space $\mathscr{X}$ is the hypercube $[0,1]^d$, with $d \in \{4, 6, 8\}$, and we consider designs $\mathbf{X}_n$ given by the first n points of a scrambled Sobol' sequence in $\mathscr{X}$, with $n \in \{10\,d, 20\,d, 50\,d, 100\,d, 200\,d\}$. The measure μ is uniform on set $\mathscr{X}_N$ given by the first $N = 2^{13+\lfloor d/2 \rfloor}$ Sobol' points in $\mathscr{X}$.

We suppose that the data are generated with the model $\mathsf{GP}(0, \sigma^2 K)$ where $\sigma^2 = 1$ and $K(\mathbf{x}, \mathbf{x}') = \psi_{3/2,2}(\|\mathbf{x} - \mathbf{x}'\|)$, see (5.1). The predictor η_n is the BLUP η_n^* for the kernel $K^{(p)}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_p}(\|\mathbf{x} - \mathbf{x}'\|)$, see (B.1) and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ assume the model $\mathsf{GP}(0, \sigma_e^2, K^{(e)})$ with $K^{(e)}(\mathbf{x}, \mathbf{x}') = \psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}'\|)$, see (B.2). As we do not simulate data (we calculate exact average performance), we cannot estimate θ_p and θ_{BLP} from $\mathbf{y}_n$. We thus adapt their choice to the design, following the suggestion in [3]: θ_p (respectively, θ_{BLP}) is such that $\psi_{5/2,\theta_p}(D_n) = 0.25$ (respectively, $\psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(D_n) = 0.25$), with $D_n = D_n[k]$ the largest of the distances from the N point in $\mathscr{X}_N$ to their k-th nearest neighbor in $\mathbf{X}_n$. It ensures that for every point $\mathbf{x}$ in $\mathscr{X}_N$ there exist at least k points $\mathbf{x}_i$ in $\mathbf{X}_n$ such that $\psi_{5/2,\theta_p}(\|\mathbf{x} - \mathbf{x}_i\|) \geq 0.25$ (respectively, $\psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}_i\|) \geq 0.25$). This implies that we assume more regularity for smaller designs: it is indeed illusory to pretend to model a highly variable function if $\mathbf{X}_n$ is very sparse (see Section G for an illustration). The choice of k is not critical and we use $k = 5$.

The left-hand side of Figure 12 shows that $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ is slightly closer than $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ to $\mathsf{E}\{\mathsf{ISE}(\eta_n)\}$ for $n = 10\,d$ and $n = 20\,d$, but the difference is not visible for larger n. On the right-hand side the plots of $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ and $\mathsf{MSE}\{\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)\}$ are practically confounded. (As the calculation of $\mathsf{var}\{\mathsf{ISE}(\eta_n)\}$ requires the computation of double integral (a double sum in this example), see (3.11) and (3.14), we omit the term $\mathsf{var}\{\mathsf{ISE}(\eta_n)\}$ is the calculation of $\mathsf{MSE}\{\widehat{\mathsf{ISE}}(\eta_n)\}$, see (3.9): our plots of $\mathsf{MSE}\{\widehat{\mathsf{ISE}}(\eta_n)\}$ thus present in fact $\mathsf{MSE}\{\widehat{\mathsf{ISE}}(\eta_n)\} - \mathsf{var}\{\mathsf{ISE}(\eta_n)\}$.) This suggests that there is little point in using $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ rather than the simpler estimator $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$. $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ overestimates $\mathsf{ISE}(\eta_n)$ is all the cases considered.

To show the importance of an appropriate choice for the kernel $K^{(e)}$, we keep the same K and $K^{(p)}$ as before (with θ_p thus adapted to the design via the rule $\psi_{5/2,\theta_p}(D_n[5]) = 0.25$) but use a fixed θ_{BLP} independently of $\mathbf{X}_n$. We first set $\theta_{\mathrm{BLP}} = 1$ (top row of Figure 13): the model used is not flexible enough and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ severely overestimate $\mathsf{ISE}(\eta_n)$ ($\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ is the same as in the top row of Figure 12). When $\theta_{\mathrm{BLP}} = 20$ (second row of Figure 13), performance deteriorates compared to Figure 13-top but is similar to that of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$. A further increase in θ_{BLP} leads to the independent limit behavior studied in Section 4.5, here with slightly poorer performance than $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$. A rather general observation is that a value $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ larger than $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ indicates a bad choice of $K^{(e)}$.

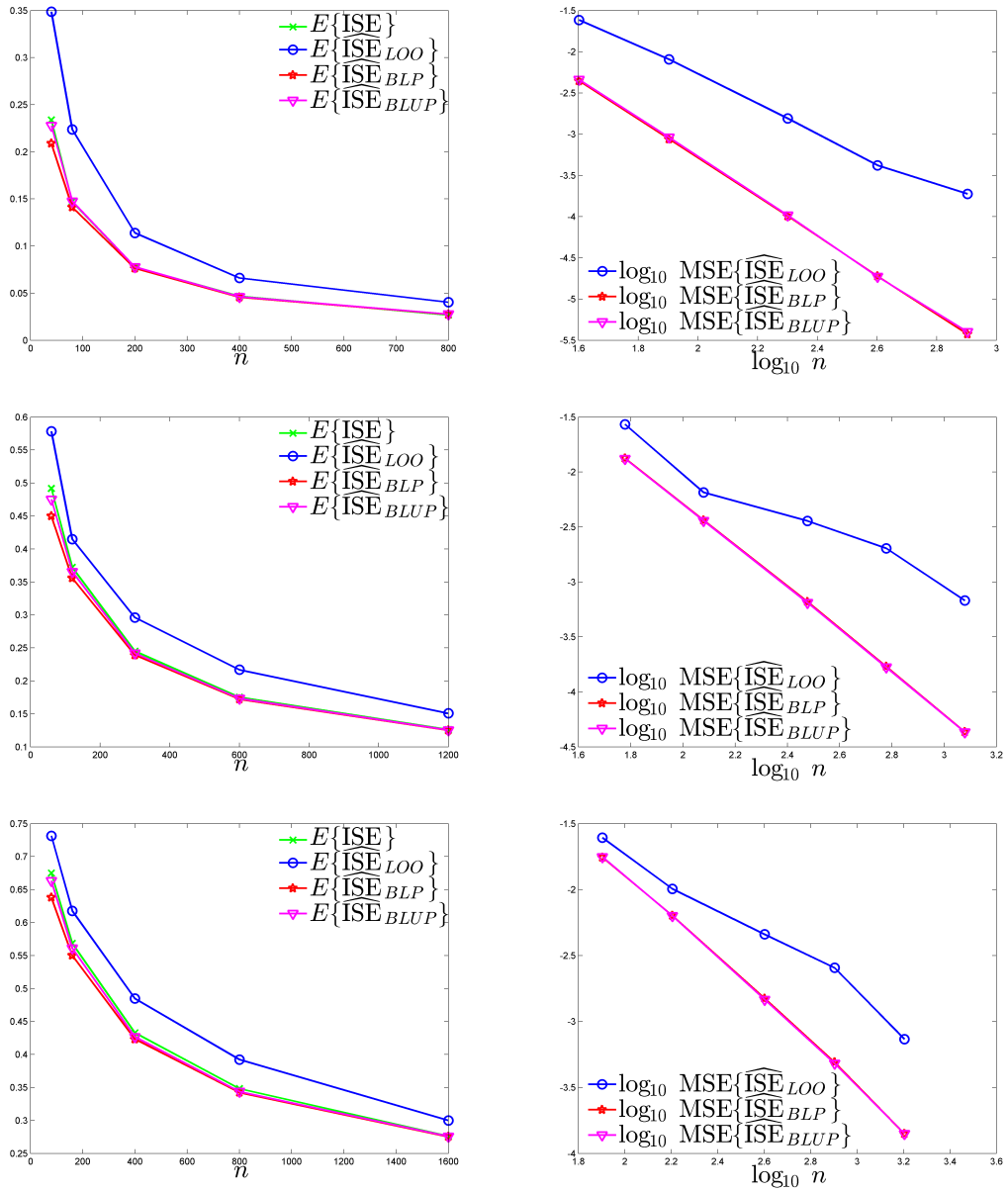

Figure 12: Performance of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ when $Y_{\mathbf{x}} \sim \mathsf{GP}(0, K_{3/2,2})$; η_n is the BLUP for $K^{(p)}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_p}(\|\mathbf{x} - \mathbf{x}'\|)$; $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ use the kernel $K^{(e)}(\mathbf{x}, \mathbf{x}') = \psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}'\|)$. From top to bottom: $d = 4, 6, 8$.

D Behavior for random functions that are not GP realizations

The computation of $\mathsf{ISE}(\eta_n)$ requires the evaluation of f on a large set of points $\mathscr{X}_N$ (we have used $N = 2^{10}$ Sobol' points in Sections 5.2 and B), which is restrictive if we want to generate f as the realization of a GP (we need to manipulate $N \times N$ matrices). In this section we follow a different route and (i) simulate a GP on a set $\mathbf{Z}_m$ of small size m, then (ii) construct f_m as the BLUP, for another GP model, on the design $\mathbf{Z}_m$. The evaluation of f_m on $\mathscr{X}_N$ then only

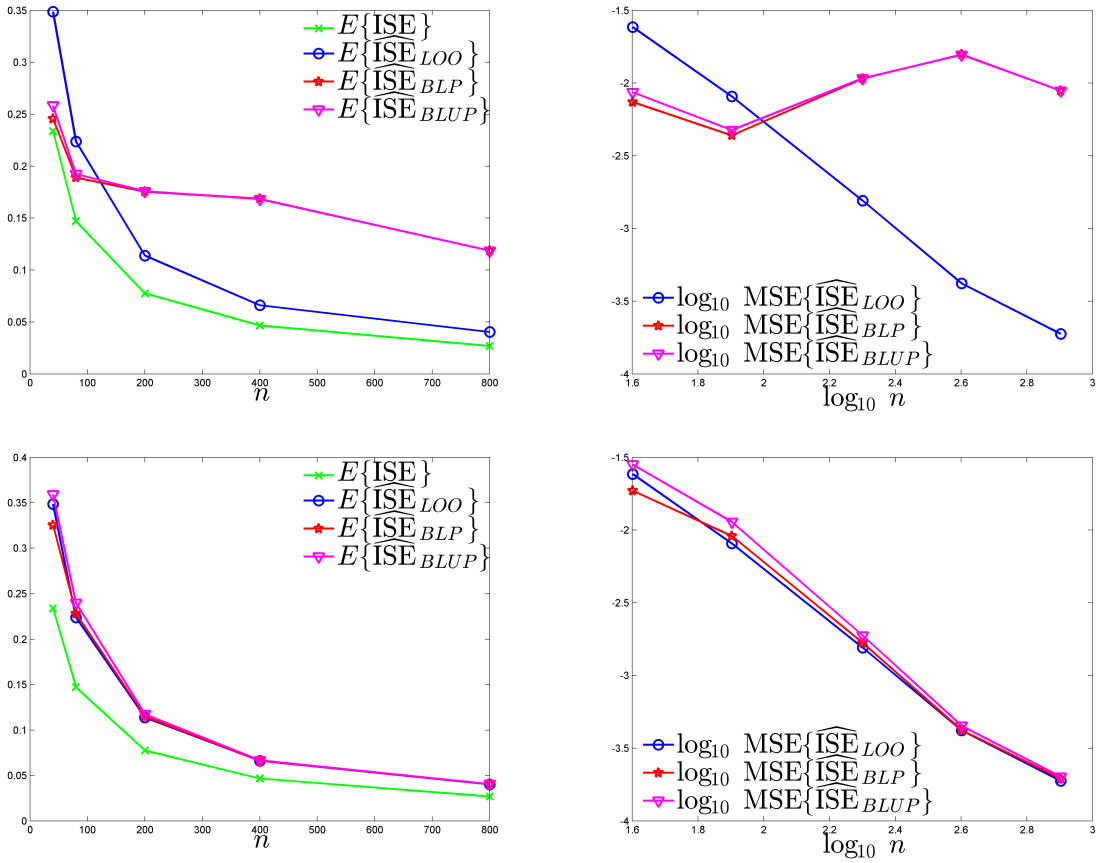

Figure 13: Same as Figure 12 but with $\theta_{\mathrm{BLP}} = 1$ (top row) and $\theta_{\mathrm{BLP}} = 20$ (second row): $\widehat{\mathsf{ISE}}_{BLP}(\eta_n) > \widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ is a sign of a poor choice of θ_{BLP}.

involves matrices of size $m \times N$. As Figure 14 illustrates in the case $d = 1$, the complexity of the functions f_m can be controlled by the value of m.

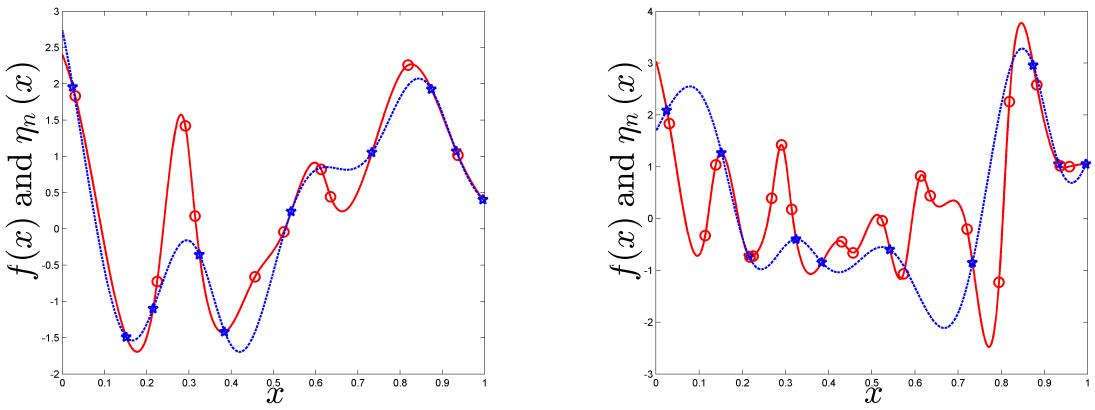

Figure 14: One realization of a random function f_m (—) given by the BLUP for a GP simulation at the design points $\mathbf{Z}_m$ ($\circ$); prediction η_n of f_m ($\cdots$) based on evaluations of f_m at the design points $\mathbf{X}_n$ ($\star$). Left: $m = n = 10$; right: $n = 10$, $m = 20$ ($\mathbf{X}_n$ is identical on both sides).

D.1 Simulations with various d and n

The design space is always the hypercube $\mathscr{X} = [0,1]^d$ and the design $\mathbf{X}_n$ is given by the first n points of a scrambled Sobol' sequence in $\mathscr{X}$. As in Section C, the measure μ is uniform on set $\mathscr{X}_N$ given by the first $N = 2^{13+\lfloor d/2 \rfloor}$ Sobol' points in $\mathscr{X}$. The design $\mathbf{Z}_m$ corresponds to the first m points of another scrambled Sobol' sequence in $\mathscr{X}$ ($\mathbf{Z}_m$ is changed for each simulation of a random function). Data simulation on $\mathbf{Z}_m$ is with the model $\mathsf{GP}(0, \sigma^2 K)$ where $\sigma^2 = 1$ and $K(\mathbf{x}, \mathbf{x}') = \psi_{3/2,50}(\|\mathbf{x} - \mathbf{x}'\|)$, see (5.1); f_m is the BLUP for $K(\mathbf{x}, \mathbf{x}') = \psi_{3/2,\theta^0}(\|\mathbf{x} - \mathbf{x}'\|)$ based on the data generated on $\mathbf{Z}_m$. As in Section C, the predictor η_n whose ISE we want to estimate is the BLUP for $K^{(p)}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_p}(\|\mathbf{x} - \mathbf{x}'\|)$, see (B.1), $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ assume the model $\mathsf{GP}(0, \sigma_e^2, K^{(e)})$ with $K^{(e)}(\mathbf{x}, \mathbf{x}') = \psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}'\|)$, see (B.2). The value of θ^0 is chosen as in Section C and satisfies $\psi_{3/2,\theta^0}(D_n[k]) = 0.25$ (with $k = 5$); θ_p and θ_{BLP} are given by the LOO estimates $\widehat{\theta}_{LOO}$ for the corresponding models: θ_p minimizes $\widehat{\mathsf{ISE}}_{LOO}(\eta_n^*)$ for the BLUP η_n^* associated with the model $\mathsf{GP}(0, \sigma^2 K_{5/2,\theta})$ and θ_{BLP} does the same for the model $\mathsf{GP}(0, \sigma^2 K_{\mathsf{IM},\theta})$. Figure 14 gives an illustration for $d = 1$ and shows a realization of f_m with the predictor η_n for $n = 10$ with $m = 10$ (left) and $m = 20$ (right).

Figure 15 presents boxplots of $\mathsf{ISE}(\eta_n)$, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ for different d and (small) designs of size $n = 10\,d$ (see [1]), obtained from 100 realizations of random f_m generated as indicated above, with $m = n$. Therefore, $m = 10\,d$, and the construction used makes f_m easier to approximate as d increases, hence the observation of decreasing values of $\mathsf{ISE}(\eta_n)$ with d. For all values of d considered, estimation of $\mathsf{ISE}(\eta_n)$ is more precise with $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ (both behave similarly) than with $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, but this superiority tends to vanish as d increases. The complexity of the evaluation of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is of the

order $\mathcal{O}(Nn^3)$, see Section 4.1, and it grows similarly for $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$. For $d = 4$, $n = 200$ and $N = 2^{15}$, the average computational time[4] for the joint evaluations of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ is about 0.7 s (for 100 repetitions, with standard deviation $\simeq 0.017$).

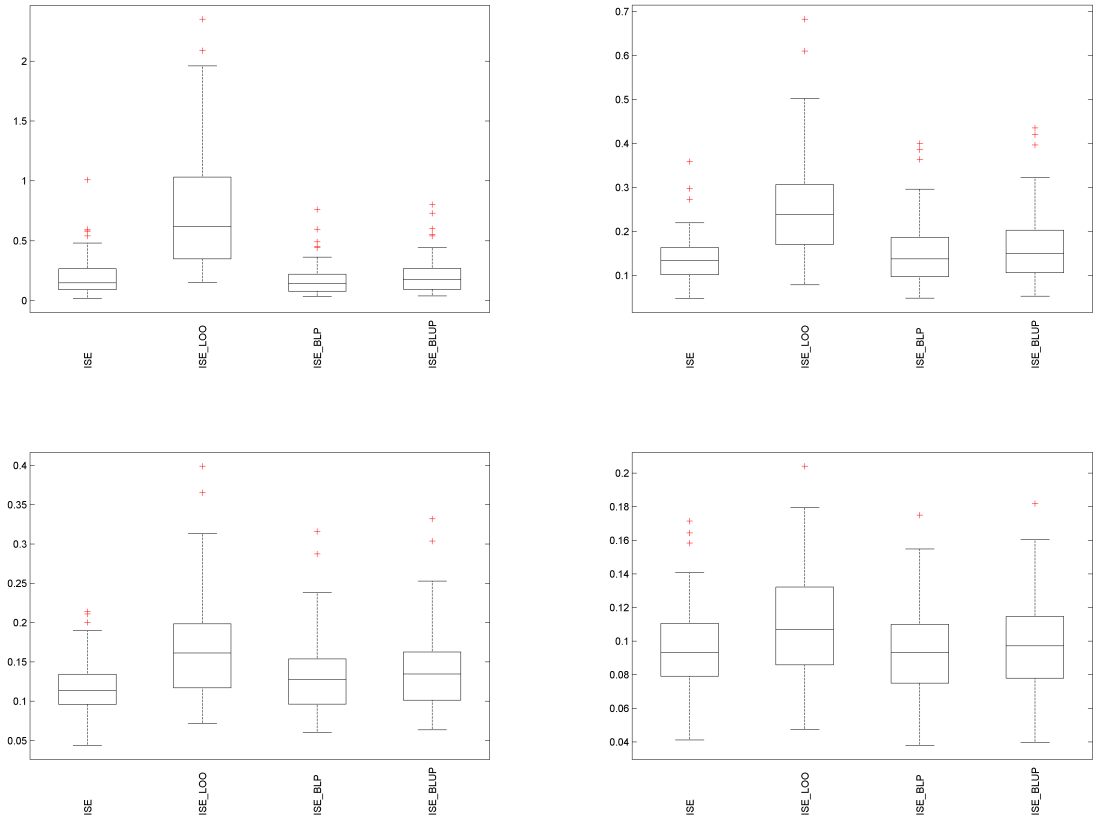

Figure 15: Boxplots of $\mathsf{ISE}(\eta_n)$, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ for random functions f_m (100 realizations) and Sobol' designs $\mathbf{X}_n$ with $m = n = 10\,d$. From top to bottom and left to right: $d = 2, 4, 6, 8$.

We take now $m = 5\,n$, making the functions f_m much more complex than above where we had $m = n$. Let us consider the case $d = 4$ (with still $n = 10\,d$). The same predictor η_n (i.e., the BLUP for $K_{5/2,\theta_p}$ with $\theta_p = \widehat{\theta}_{LOO}$) now performs very poorly: compare the boxplots of $\mathsf{ISE}(\eta_n)$ on the left panel of Figure 16 and on the top-right panel of Figure 15. In fact, η_n performs even worse than the simple empirical mean (i.e., $\overline{\eta_n} = \mathbf{1}_n^\top \mathbf{y}_n/n$), whose performance is shown on the right panel of Figure 16. The three ISE estimators $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ are capable to uncover this poor performance of η_n.

[4]Computations are in Matlab, on a PC with a clock speed of 2.5 GHz and 32 GB RAM.

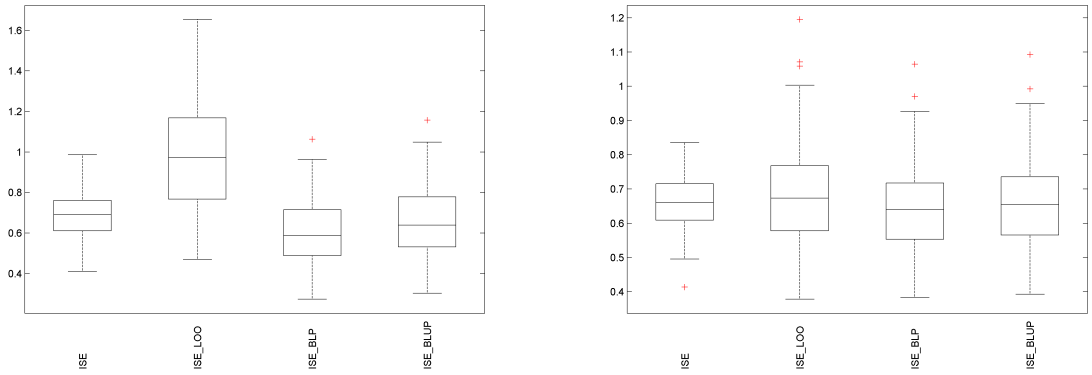

Figure 16: Same as Figure 15-top-right but for random functions f_m with $m = 200$ and a Sobol' design $\mathbf{X}_n$ with $n = 40$ in $[0,1]^4$. Left: η_n is the BLUP for $K_{5/2,\theta_p}$ with θ_p estimated by LOOCV; right: η_n is the empirical mean $\overline{\eta_n} = \mathbf{1}_n^\top \mathbf{y}_n/n$.

On the contrary, if we keep $m = 10\,d$ and increase n, $\mathsf{ISE}(\eta_n)$ decreases and is difficult to estimate accurately (in terms of relative precision). Figure 17 is for $d = 4$, $m = 40$ and $n = 400$. On the left panel, $\theta_{\text{BLP}} = \widehat{\theta}_{LOO}$ (which gives $\theta_{\text{BLP}} \in (1.55, 2.25)$ with an average value $\simeq 1.94$ for the 100 realizations); on the right panel, θ_{BLP} is chosen with the rule of Section C, i.e., $\psi_{\mathsf{IM},\theta_{\text{BLP}}}(D_n[5]) = 0.25$ (which gives $\theta_{\text{BLP}} \simeq 3.8$). We can notice slightly better performance for $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ on the right-hand panel, but the main observation concerns the low sensitivity to the choice of $K^{(e)}$ and the relevance of the rule of Section C (for which, moreover, no numerical optimization with respect to θ_{BLP} is required).

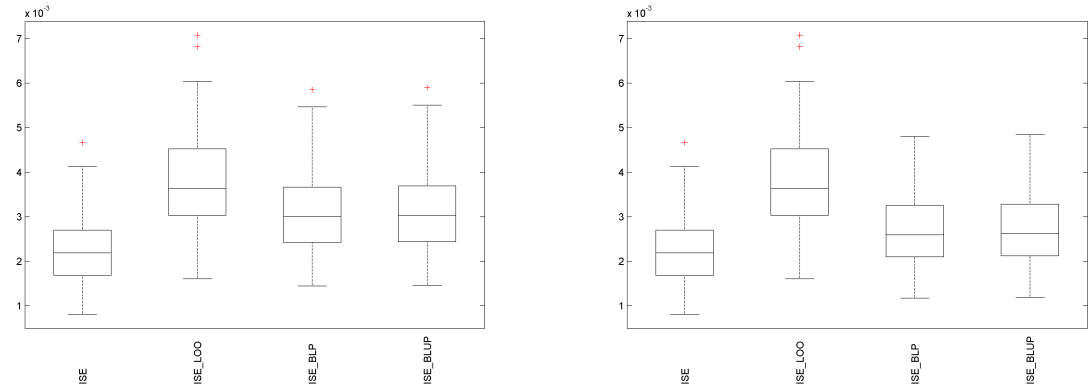

Figure 17: Same as Figure 15-top-right (random functions f_m with $m = 40$) but with $\mathbf{X}_n$ a Sobol' design with $n = 400$ points in $[0,1]^4$. Left: $\theta_{\text{BLP}} = \widehat{\theta}_{LOO}$ (and $\theta_{\text{BLP}} \in (1.55, 2.25)$); right: $\theta_{\text{BLP}} \simeq 3.8$ satisfies $\psi_{\mathsf{IM},\theta_{\text{BLP}}}(D_n[5]) = 0.25$. $\mathsf{ISE}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ are the same on both panels.

D.2 Noisy observations

We consider the same experimental framework as in Section D.1 when $d = 4$, with $m = n = 10\,d = 40$, but the observations $\mathbf{y}_n$ are now given by $y(\mathbf{x}_i) = f_m(\mathbf{x}_i) + \zeta_i$, $i = 1, \ldots, n$, where the measurement errors ζ_i are i.i.d. $\mathcal{N}(0, \gamma^2)$. Following Section 4.7, for the construction of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ (and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$) we assume that the data obey the model $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2 K_{r^{(e)}}^{'(e)})$, where $K_r^{'(e)}(\mathbf{x}, \mathbf{x}') = K^{(e)}(\mathbf{x}, \mathbf{x}') + r\,\delta_{\mathbf{x},\mathbf{x}'} = \psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}'\|) + r\,\delta_{\mathbf{x},\mathbf{x}'}$ with $\delta_{\mathbf{x},\mathbf{x}'} = 1$ when $\mathbf{x} = \mathbf{x}'$ and is zero otherwise. We do not attempt to estimate $r^{(e)}$ from the data, but rather investigate the dependence of performance on the choice of $r^{(e)}$. The construction of f_m is based on GP's with variance $\sigma^2 = 1$ (see Section D.1), and we have set a fairly high noise level $\gamma = 0.25$. The predictor η_n is now the BLUP for the kernel $K^{(p)}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_p}(\|\mathbf{x} - \mathbf{x}'\|) + \gamma^2\,\delta_{\mathbf{x},\mathbf{x}'}$; θ_p (respectively, θ_{BLP}) is obtained by minimization of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n^*)$, with η_n^* the BLUP for $\mathsf{GP}(0, K^{(p)})$ (respectively, for $\mathsf{GP}(0, K_r^{'(e)})$).

Figure 18 shows boxplots of $\mathsf{ISE}(\eta_n)$, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$ for 100 random functions f_m and noise realizations, for four different $r^{(e)}$: $r^{(e)} = \gamma^2 = 0.0625$ (top left), which can be considered as a natural choice in the present context; a severely underestimated value $r^{(e)} = \gamma^2/10$ (top right); and two overestimated values, $r^{(e)} = 5\,\gamma^2$ and $r^{(e)} = 10\,\gamma^2$ (second row). Unsurprisingly, the best performance is obtained for $r^{(e)}$ close to γ^2, but using a (much) smaller value has little effect; performance deteriorates when $r^{(e)}$ becomes much larger than γ^2, but remains acceptable (and superior to that of $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$).

D.3 Unreliable estimation of quantiles and conditional values-at-risk

The method proposed in the paper provides an estimate $\widehat{\varepsilon_n^2}(\mathbf{x})$ of $\varepsilon_n^2(\mathbf{x})$ at any $\mathbf{x}$, and in all the examples presented, $\mathsf{ISE}(\eta_n)$ has been estimated by the empirical mean $(1/N)\sum_{i=1}^N \widehat{\varepsilon_n^2}(\mathbf{x}^{(i)})$ calculated for N points $\mathbf{x}^{(i)}$ distributed with μ. One may thus think of using the N estimates $\widehat{\varepsilon_n^2}(\mathbf{x}^{(i)})$ to compute an empirical quantile (or value-at-risk) Q_α and conditional value at risk CVaR_α (see, e.g. [4] and the references therein) at a given level α. However, the errors $\varepsilon_n^2(\mathbf{x})$ as well as the estimates $\widehat{\varepsilon_n^2}(\mathbf{x})$ are correlated[5], and the distributions of $\varepsilon_n^2(\mathbf{x})$ and $\widehat{\varepsilon_n^2}(\mathbf{x})$ may significantly differ. The following example provides an illustration.

The framework is as in Section D.1 for $d = 4$, with $m = n = 10\,d = 40$ and $N = 2^{15}$. The left panel of Figure 19 presents a scatter plot of $(\widehat{\varepsilon_{n\,BLP}^2}(\mathbf{x}^{(i)}), \varepsilon_n^2(\mathbf{x}^{(i)}))$ for one random f_m, with a red solid line showing the first diagonal. There are more small squared errors $\varepsilon_n^2(\mathbf{x}^{(i)})$ than small squared errors $\widehat{\varepsilon_{n\,BLP}^2}(\mathbf{x}^{(i)})$, but some $\varepsilon_n^2(\mathbf{x}^{(i)})$ are much larger than $\widehat{\varepsilon_{n\,BLP}^2}(\mathbf{x}^{(i)})$. The right panel shows, for the same simulation, the empirical c.d.f. F_T of the 2^{15} true squared prediction errors $\varepsilon_n^2(\mathbf{x}^{(i)})$ (red solid line) and the empirical c.d.f. F_{BLP} and F_{BLUP} of the BLP and BLUP estimates $\widehat{\varepsilon_{n\,BLP}^2}(\mathbf{x}^{(i)})$ and $\widehat{\varepsilon_{n\,BLUP}^2}(\mathbf{x}^{(i)})$ of Sections 4.1 and 4.6, respectively in blue solid line and green dashed line. The behavior observed is typical: $F_{BLP}(t)$ and $F_{BLUP}(t)$ are very close; $F_T(t)$ is larger than $F_{BLP}(t)$ and $F_{BLUP}(t)$ for small t but is smaller for large t as $\widehat{\varepsilon_{n\,BLP}^2}(\mathbf{x})$ and $\widehat{\varepsilon_{n\,BLUP}^2}(\mathbf{x})$ tend to smooth $\varepsilon_n^2(\mathbf{x})$.

[5] With the notation of Section 3.1, under the assumption $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma^2\,K)$, $\mathsf{cov}\{\varepsilon_n^2(\mathbf{x}), \varepsilon_n^2(\mathbf{x}')\} = 2\,\sigma^4\,\rho_n^4(\mathbf{x}, \mathbf{x}')$ and $\mathsf{cov}\{\widehat{\varepsilon_n^2}(\mathbf{x}), \widehat{\varepsilon_n^2}(\mathbf{x}')\} = 2\,\sigma^4\,\boldsymbol{\beta}^\top(\mathbf{x})(\mathbf{R}_n^\top \mathbf{K}_n \mathbf{R}_n)^{\odot 2}\boldsymbol{\beta}(\mathbf{x}')$ with $\boldsymbol{\beta}(\mathbf{x})$ given by (4.1) for $\widehat{\varepsilon_{n\,BLP}^2}(\mathbf{x})$ (Section 4.1) and by (4.8) for the unbiased version $\widehat{\varepsilon_{n\,BLUP}^2}(\mathbf{x})$ (Section 4.6).

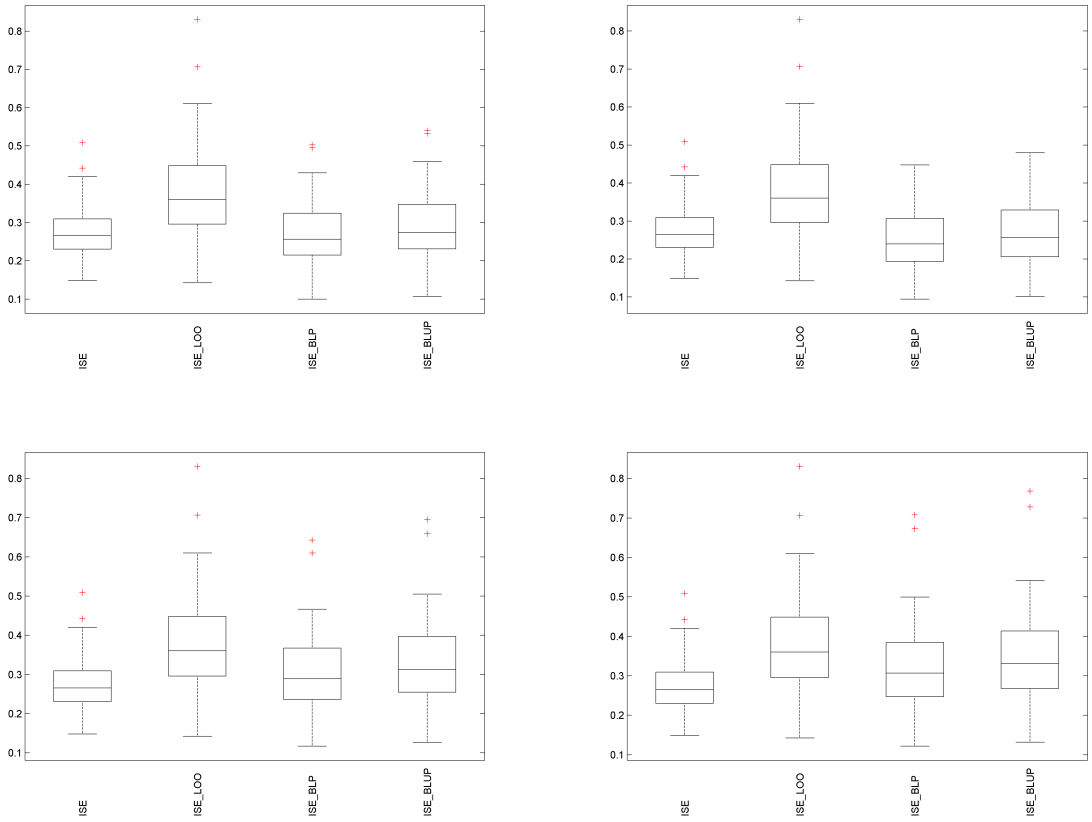

Figure 18: Same as Figure 15-top-right (random functions f_m with $d = 4$ and $m = n = 40$) but with noisy observations with standard deviation $\gamma = 0.25$. First row: $r^{(e)} = \gamma^2$ (left) and $r^{(e)} = 0.1\,\gamma^2$ (right); second row: $r^{(e)} = 5\,\gamma^2$ (left) and $r^{(e)} = 10\,\gamma^2$ (right). $\mathsf{ISE}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ are the same on all panels.

Figure 20 shows boxplots of Q_α (left) CVaR_α (right) for the true squared prediction errors $\varepsilon_n^2(\mathbf{x}^{(i)})$ and the BLP and BLUP estimates $\widehat{\varepsilon}_{nBLP}^2(\mathbf{x}^{(i)})$ and $\widehat{\varepsilon}_{nBLUP}^2(\mathbf{x}^{(i)})$ for $\alpha = 0.95$ (top row) and $\alpha = 0.5$ (bottom row), for 100 random functions f_m. In agreement with Figure 19-right, we observe that Q_α is strongly underestimated (respectively, overestimated) for $\alpha = 0.95$ (respectively, $\alpha = 0.5$). The presence of squared errors $\varepsilon_n^2(\mathbf{x}^{(i)})$ much larger than $\widehat{\varepsilon}_{nBLP}^2(\mathbf{x}^{(i)})$ explains that CVaR_α is underestimated for both values of α (however, performance improves when α decreases, as $\mathsf{CVaR}_\alpha \to \mathsf{ISE}(\eta_n)$ when $\alpha \to 0$). The information that the $\widehat{\varepsilon}_{nBLP}^2(\mathbf{x}^{(i)})$ and $\widehat{\varepsilon}_{nBLUP}^2(\mathbf{x}^{(i)})$ provide on the tail distribution of the squared prediction errors $\varepsilon_n^2(\mathbf{x})$ is therefore very unreliable. We have observed similar disappointing behavior with other examples.

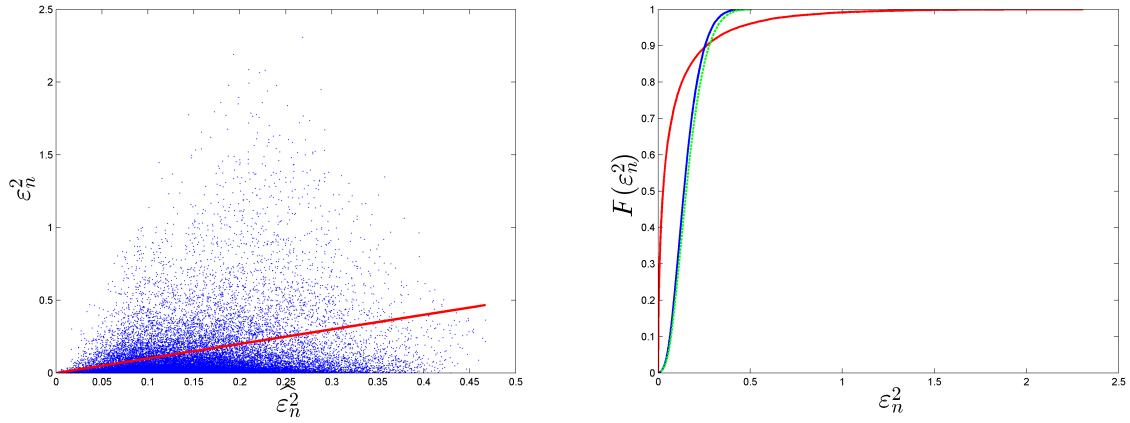

Figure 19: Left: scatter plot of $(\widehat{\varepsilon^2_{n\,BLP}}(\mathbf{x}^{(i)}), \varepsilon_n^2(\mathbf{x}^{(i)}))$ for one random f_m. Right: empirical c.d.f. of the true squared errors $\varepsilon_n^2(\mathbf{x}^{(i)})$ (—) and of their BLP and BLUP estimates $\widehat{\varepsilon^2_{n\,BLP}}(\mathbf{x}^{(i)})$ (—) and $\widehat{\varepsilon^2_{n\,BLUP}}(\mathbf{x}^{(i)})$ (···), for the same random f_m.

E GP with parameterized mean

The developments of Section 4 assumed that f is the realization of GP with zero mean. Here we show how to estimate the ISE of a given linear predictor $\eta_n(\cdot) = \mathbf{w}_n^\top(\cdot)\mathbf{y}_n$ when this assumption is relaxed.

E.1 Universal kriging model

Consider the framework of universal kriging, and assume that f is the realization of a GP $Y_\mathbf{x} \sim \mathsf{GP}(\boldsymbol{\tau}^\top \mathbf{h}(\mathbf{x}), \sigma^2 K)$, with $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_p(\mathbf{x})]^\top$ a vector of p known functions on $\mathscr{X}$ and $\boldsymbol{\tau}$ a vector of unknown parameters. We then have

$$\begin{aligned} \mathsf{ISE}(\eta_n) &= \int_\mathscr{X} [Y_\mathbf{x} - \mathbf{w}_n^\top(\mathbf{x})\mathbf{y}_n]^2 \, \mu(\mathrm{d}\mathbf{x}) \\ &= \int_\mathscr{X} \left[Z_\mathbf{x} + \boldsymbol{\tau}^\top \mathbf{h}(\mathbf{x}) - \mathbf{w}_n^\top(\mathbf{x})(\mathbf{z}_n + \mathbf{H}_n\boldsymbol{\tau}) \right]^2 \mu(\mathrm{d}\mathbf{x}) \,, \end{aligned}$$

where $Z_\mathbf{x} = Y_\mathbf{x} - \boldsymbol{\tau}^\top \mathbf{h}(\mathbf{x}) \sim \mathsf{GP}(0, \sigma^2 K)$, $\mathbf{H}_n$ is the $n \times p$ matrix with i-th row equal to $\mathbf{h}^\top(\mathbf{x}_i)$, and $\mathbf{z}_n = \mathbf{y}_n - \mathbf{H}_n\boldsymbol{\tau}$. This gives

$$\mathsf{ISE}(\eta_n) = \mathsf{ISE}_0(\eta_n) + I(\boldsymbol{\tau}) + 2\, I_n(\boldsymbol{\tau}) \,, \tag{E.1}$$

where

$$\begin{aligned} \mathsf{ISE}_0(\eta_n) &= \int_\mathscr{X} [Z_\mathbf{x} - \mathbf{w}_n^\top(\mathbf{x})\mathbf{z}_n]^2 \, \mu(\mathrm{d}\mathbf{x}) \,, \qquad (\text{E.2}) \\ I(\boldsymbol{\tau}) &= \int_\mathscr{X} \left\{ \boldsymbol{\tau}^\top \left[\mathbf{h}(\mathbf{x}) - \mathbf{H}_n^\top \mathbf{w}_n(\mathbf{x}) \right] \right\}^2 \mu(\mathrm{d}\mathbf{x}) \,, \\ I_n(\boldsymbol{\tau}) &= \int_\mathscr{X} [Z_\mathbf{x} - \mathbf{w}_n^\top(\mathbf{x})\mathbf{z}_n] \left\{ \boldsymbol{\tau}^\top \left[\mathbf{h}(\mathbf{x}) - \mathbf{H}_n^\top \mathbf{w}_n(\mathbf{x}) \right] \right\} \mu(\mathrm{d}\mathbf{x}) \,. \end{aligned}$$

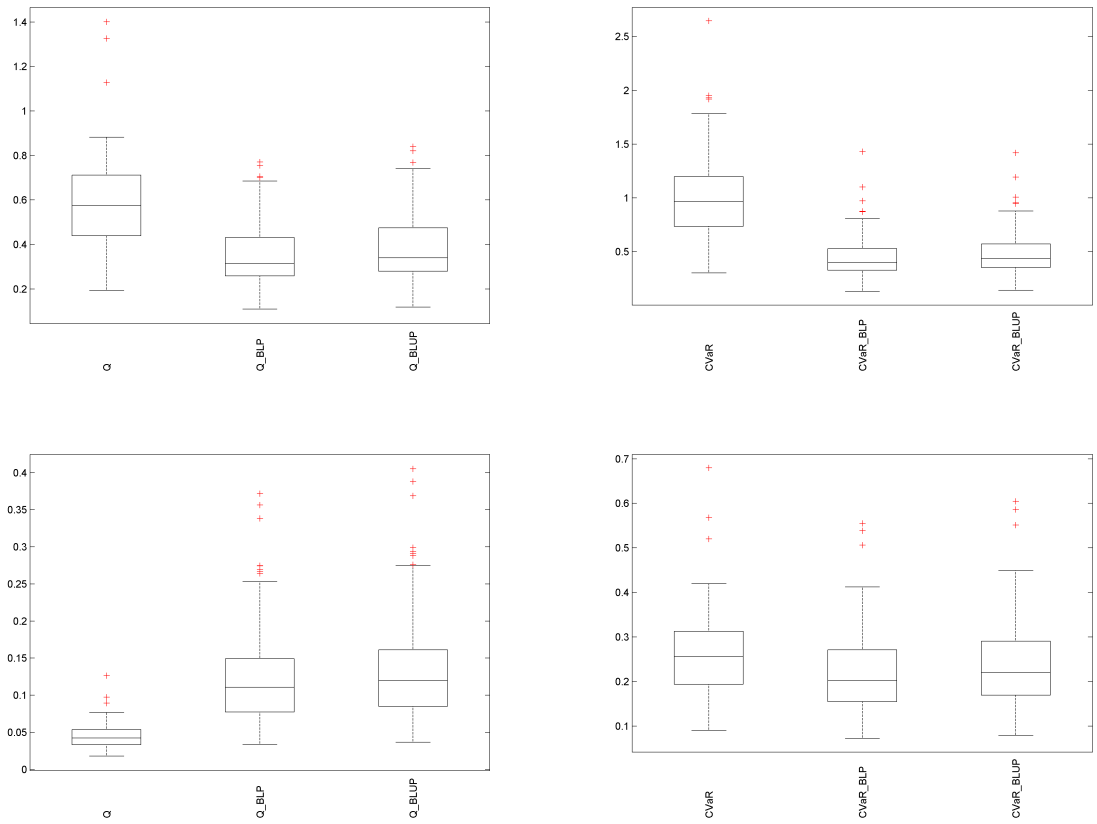

Figure 20: Boxplots of the quantiles Q_α (left) conditional value-at-risk CVaR_α (right) for the $\varepsilon_n^2(\mathbf{x}^{(i)})$ and their estimates $\widehat{\varepsilon_n^2}(\mathbf{x}^{(i)})$ (the $\varepsilon_n^2(\mathbf{x}^{(i)})$ and $\widehat{\varepsilon_n^2}(\mathbf{x}^{(i)})$ are the same as in the top-right panel of Figure 15); top row: $\alpha = 0.95$; bottom row: $\alpha = 0.5$.

In (E.1), $\mathsf{ISE}_0(\eta_n)$ is the ISE for the centered GP model $Z_\mathbf{x} \sim \mathsf{GP}(0, \sigma^2 K)$, which can be estimated with the method presented in Section 4, $I(\boldsymbol{\tau})$ is a constant and $I_n(\boldsymbol{\tau})$ has zero mean.

A simple approach to estimate $\mathsf{ISE}(\eta_n)$, assuming a kernel $K^{(e)}$, is therefore as follows.

(i) Estimate the parametric trend, using for example the BLUE for $\boldsymbol{\tau}$ given by

$$\widehat{\boldsymbol{\tau}}^n = (\mathbf{H}_n^\top {\mathbf{K}_n^{(e)}}^{-1} \mathbf{H}_n)^{-1} \mathbf{H}_n^\top {\mathbf{K}_n^{(e)}}^{-1} \mathbf{y}_n \,. \tag{E.3}$$

(ii) Remove $\mathbf{H}_n \widehat{\boldsymbol{\tau}}^n$ from the observations $\mathbf{y}_n$ and estimate $\mathsf{ISE}_0(\eta_n)$ (E.2) for these centered observations $\mathbf{z}_n$ under the assumption $Z_\mathbf{x} \sim \mathsf{GP}(0, \sigma_e^2 K^{(e)})$;

(iii) Add $I(\widehat{\boldsymbol{\tau}}^n)$ to the estimated ISE.

This approach neglects the error due to the estimation of $\boldsymbol{\tau}$, which is acceptable when $p \ll n$; other approaches, more accurate, could certainly be developed, at the expense of increased complexity. A direct application of the approach of Section 4.1 through the calculation of $\mathsf{E}\{\varepsilon^2(\mathbf{x})\varepsilon_{-i}^2\}$ and $\mathsf{E}\{\varepsilon_{-i}^2(\mathbf{x})\varepsilon_{-j}^2\}$ for the model $Y_\mathbf{x} \sim \mathsf{GP}(\boldsymbol{\tau}^\top \mathbf{h}(\mathbf{x}), \sigma^2 K)$ would also be possible. However, these expressions depend explicitly on $\boldsymbol{\tau}$ and σ^2, whereas the estimation of $\mathsf{ISE}_0(\eta_n)$ by (4.2) does not require the construction of an estimator of σ^2.

E.2 Ordinary kriging model

The model $Y_{\mathbf{x}} \sim \mathsf{GP}(\tau, \sigma^2 K)$ with $\tau \in \mathbb{R}$ and $\mathbf{h}(\mathbf{x}) \equiv 1$ (ordinary kriging) is frequently used. In this case we get $\widehat{\tau}^n = (\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{y}_n)/(\mathbf{1}_n^\top \mathbf{K}_n^{-1} \mathbf{1}_n)$ and $I(\tau) = \tau^2 \int_{\mathscr{X}} [1 - \mathbf{w}_n^\top(\mathbf{x}) \mathbf{1}_n]^2 \, \mu(\mathrm{d}\mathbf{x})$.

If η_n is such that $\mathbf{w}_n^\top(\mathbf{x}) \mathbf{1}_n = 1$ for all $\mathbf{x}$, then any translation of the observations leaves the ISE invariant (since the prediction η_n itself is invariant) and the LOO residuals ε_{-i} are invariant too. We have $\mathsf{ISE}(\eta_n) = \int_{\mathscr{X}} [Z_{\mathbf{x}} - \mathbf{w}_n^\top(\mathbf{x}) \mathbf{z}_n]^2 \, \mu(\mathrm{d}\mathbf{x})$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ can be calculated for a centered $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$ without centering the observations. The case of the ordinary kriging predictor is a typical example. More generally, denote by $\mathsf{E}_\tau\{\cdot\}$ and $\mathsf{MSE}_\tau\{\cdot\}$ the expectation and MSE under the model $\mathsf{GP}(\tau, \sigma^2 K)$. Direct calculation shows that, when $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ is calculated for $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$ (i.e., assuming that $\tau = 0$), we have

$$\mathsf{E}_\tau\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} = \mathsf{E}_0\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} + \tau^2 \, (\mathbf{1}_n^\top \mathbf{R}_n)^{\odot 2} \mathbf{S}_n^{-1} \mathbf{b}_n \,,$$

where $\mathsf{E}_0\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ is given by (4.5), $\mathbf{R}_n$ is the matrix in (3.1) and $\mathbf{S}_n$ and $\mathbf{b}_n$ are respectively given by (3.5) and (3.12). We obtain similarly

$$\mathsf{MSE}_\tau\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} = \mathsf{MSE}_0\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\} + \tau^2 \, C_n \,,$$

where $\mathsf{MSE}_0\{\widehat{\mathsf{ISE}}_{BLP}(\eta_n)\}$ is given by (4.6) and C_n tends to zero when $\int_{\mathscr{X}} [1 - \mathbf{w}_n^\top(\mathbf{x}) \mathbf{1}_n]^2 \, \mu(\mathrm{d}\mathbf{x})$ and $\|\mathbf{R}_n^\top \mathbf{1}_n\|$ tend to zero. This indicates that the performance of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ constructed under the assumption $\tau = 0$ is preserved when τ is small or when the predictor η_n is such that, for all n and $\mathbf{X}_n$, $\mathbf{w}_n^\top(\mathbf{x}) \mathbf{1}_n \approx 1$ for all $\mathbf{x}$; see Section 5.3 for an example. Similar developments show that $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ calculated for $\mathsf{GP}(0, \sigma_e^2 K^{(e)})$ behaves similarly when the true data generating model is $\mathsf{GP}(\boldsymbol{\tau}^\top \mathbf{h}(\mathbf{x}), \sigma^2 K)$ or $\mathsf{GP}(0, \sigma^2 K)$ provided that the predictor η_n satisfies $\mathbf{H}_n^\top \mathbf{w}_n(\mathbf{x}) \approx \mathbf{h}(\mathbf{x})$ for all $\mathbf{x}$. The correction of Section E.1 can be applied otherwise.

F Mixtures of GP models

In the construction of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$, instead of assuming that f is the realization of a unique GP, $Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma_e^2 K^{(e)})$, we may consider a mixture of GP; that is, consider a family $\{K_t\}_{t=1,\ldots,T}$ of T different kernels (stationary or not, with different regularities…) and assume that $Y_{\mathbf{x}} | s \sim \mathsf{GP}(0, \sigma_e^2 K_s)$, with $\mathrm{Prob}\{s = t\} = \nu_t$. (The infinite mixture model could be considered as well but is computationally more difficult to handle.) All expectations under this finite mixture model can be decomposed as

$$\mathsf{E}\{X\} = \sum_{t=1}^{T} \nu_t \mathsf{E}\{X | Y_{\mathbf{x}} \sim \mathsf{GP}(0, \sigma_e^2 K_t)\} \,.$$

This gives for instance $\mathsf{E}\{\varepsilon^2(\mathbf{x})\} = \sigma^2 \sum_{t=1}^{T} \nu_t \rho_{n,t}^2(\mathbf{x})$ where $\rho_{n,t}^2(\mathbf{x})$ is given by (3.2) with $K = K_t$, and with obvious notation $\mathsf{E}\{\boldsymbol{\varepsilon}_{LOO} \boldsymbol{\varepsilon}_{LOO}^\top\} = \sigma^2 \mathbf{R}_n^T \left(\sum_{t=1}^{T} \nu_t \mathbf{K}_{n,t} \right) \mathbf{R}_n$, $\mathsf{E}\{\boldsymbol{\varepsilon}_{LOO} \varepsilon_n(\mathbf{x})\} = \sigma^2 \mathbf{R}_n^T \sum_{t=1}^{T} \nu_t \mathbf{t}_{n,t}(\mathbf{x})$, etc. Developments similar to those of Section 4 then yield the expressions of $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$. The weights $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_T)^\top$ can be adjusted to the data $\mathbf{y}_n$, as in Bayesian Model Averaging (BMA), see [3] (with a proposition concerning the choice of prior weights in Section 5 of the same paper).

Since a mixture $\eta_n(\cdot) = \sum_{t=1}^T \nu_t\, \eta_{n,t}(\cdot)$ of kriging predictors obtained by BMA with *fixed weights* (i.e., not depending on $\mathbf{y}_n$) remains linear in $\mathbf{y}_n$, ISE estimation by $\widehat{\mathsf{ISE}}_{BLP}(\cdot)$ can also be applied to such mixture models. When the weights satisfy $\boldsymbol{\nu}^\top \mathbf{1}_T = 1$, then, with the notation of Section 3, the i-th LOO error becomes $\varepsilon_{-i} = \sum_{t=1}^T \nu_t \left[y_i - \eta_{n\setminus i,t}(\mathbf{x}_i)\right] = \sum_{t=1}^T \nu_t\, \varepsilon_{-i,t}$ and the squared LOO errors $\boldsymbol{\varepsilon}_{LOO}^{\odot 2}$ are quadratic in $\boldsymbol{\nu}$. Denoting $\mathbf{E}_{LOO}$ the $T \times n$ matrix with $\{\mathbf{E}_{LOO}\}_{t,i} = \varepsilon_{-i,t}$, any ISE estimator of the form $\widehat{\mathsf{ISE}}(\eta_n[\boldsymbol{\nu}]) = \boldsymbol{\gamma}^\top \boldsymbol{\varepsilon}_{LOO}^{\odot 2}$ (thus in particular $\widehat{\mathsf{ISE}}_{BLP}(\eta_n[\boldsymbol{\nu}])$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n[\boldsymbol{\nu}])$) can be written as

$$\widehat{\mathsf{ISE}}(\eta_n) = \boldsymbol{\nu}^\top \mathbf{E}_{LOO} \boldsymbol{\Gamma} \mathbf{E}_{LOO}^\top \boldsymbol{\nu}\,,$$

where $\boldsymbol{\Gamma} = \mathrm{diag}\{\gamma_i,\, i = 1, \ldots, n\}$. Minimization of $\widehat{\mathsf{ISE}}(\eta_n[\boldsymbol{\nu}])$ with respect to $\boldsymbol{\nu}$ under the constraint $\boldsymbol{\nu}^\top \mathbf{1}_T = 1$ yields the optimal predictor $\eta_n[\boldsymbol{\nu}^*]$ (in the sense of $\widehat{\mathsf{ISE}}(\cdot)$) with

$$\boldsymbol{\nu}^* = \frac{(\mathbf{E}_{LOO} \boldsymbol{\Gamma} \mathbf{E}_{LOO}^\top)^{-1} \mathbf{1}_T}{\mathbf{1}_T^\top (\mathbf{E}_{LOO} \boldsymbol{\Gamma} \mathbf{E}_{LOO}^\top)^{-1} \mathbf{1}_T}\,.$$

It is tempting to iterate the process, as suggested in Section 6 of the paper for model selection: indeed, the optimal weights $\boldsymbol{\nu}^*$ could be used to define a mixture of GP models for the construction of $\widehat{\mathsf{ISE}}(\eta_n[\boldsymbol{\nu}])$, whose optimization would lead to an updated optimal $\boldsymbol{\nu}^*$.

G A 1-d example of poor performance due to a bad design

This simple example illustrates a limitation of the method described in the last paragraph of the conclusion section: due to design sparsity, when the predictor is much smoother than f, the LOO squared residuals ε_{-i}^2 can be very small and $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ may severely underestimate $\mathsf{ISE}(\eta_n)$, so that the inaccuracy of η_n can remain undetected. The problem of too sparse a design relative to the variability of f is more serious in high dimensions, but this one-dimensional case already gives a picture of the possible situation.

Here $f(x) = \sum_{i=1}^5 \psi_{3/2,20}(|x - z_i|)$ with $\{z_1, \ldots, z_5\} = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and $\psi_{3/2,\theta}$ given by (5.1); $\mathbf{X}_n$ corresponds to the first n points of a scrambled Sobol' sequence in $[0, 1]$. The predictor η_n and $\widehat{\mathsf{ISE}}_{BLP}(\cdot)$ are constructed as in Section D.1, i.e., respectively with the kernels $K^{(p)}(\mathbf{x}, \mathbf{x}') = \psi_{5/2,\theta_p}(\|\mathbf{x} - \mathbf{x}'\|)$ and $K^{(e)}(\mathbf{x}, \mathbf{x}') = \psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(\|\mathbf{x} - \mathbf{x}'\|)$ given by (B.1) and (B.2); θ_p and θ_{BLP} satisfy $\psi_{5/2,\theta_p}(D_n[5]) = 0.25$ and $\psi_{\mathsf{IM},\theta_{\mathrm{BLP}}}(D_n[5]) = 0.25$, see Section C. Figure 21 shows $f(x)$ and $\eta_n(x)$, $x \in [0, 1]$, for $n = 5$ (left) and $n = 15$ (right).

In the first case, with $n = 5$ ($D_n[5] = 0.9750$, $\theta_p \simeq 1.63$ and $\theta_{\mathrm{BLP}} \simeq 1.78$), η_n is very smooth, $\max_i \varepsilon_{-i}^2 \simeq 0.0771$, $\widehat{\mathsf{ISE}}_{LOO}(\eta_n) \simeq 0.029$ and $\widehat{\mathsf{ISE}}_{BLP}(\eta_n) \simeq 3.26\,10^{-4}$, whereas $\mathsf{ISE}(\eta_n) \simeq 0.125$. At the same time, the predictor $\overline{\eta_n}$ given by the empirical mean, $\overline{\eta_n} = \mathbf{1}_n^\top \mathbf{y}_n / n$, has larger estimated ISE: $\widehat{\mathsf{ISE}}_{LOO}(\overline{\eta_n}) \simeq 0.095$ and $\widehat{\mathsf{ISE}}_{BLP}(\overline{\eta_n}) \simeq 0.061$. The inaccuracy of η_n thus remains undetected (the true ISE is $\mathsf{ISE}(\overline{\eta_n}) \simeq 0.0731$, showing that η_n is indeed worse than $\overline{\eta_n}$). It would be easily revealed by additional evaluations of f on a set of test points z_i (unless by bad luck the z_i are such that $\eta_n(z_i) \approx f(z_i)$).

The situation improves with the use of a richer design. When $n = 15$, (with $D_n[5] = 0.267$, $\theta_p \simeq 5.97$ and $\theta_{\mathrm{BLP}} \simeq 6.49$), η_n is much closer to f although $\max_i \varepsilon_{-i}^2 \simeq 0.305$ is much larger than before. We have now $\widehat{\mathsf{ISE}}_{LOO}(\eta_n) \simeq 0.1073$, $\widehat{\mathsf{ISE}}_{BLP}(\eta_n) \simeq 0.0053$ and $\mathsf{ISE}(\eta_n) \simeq 0.0097$; $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ thus underestimates $\mathsf{ISE}(\eta_n)$ by a factor of 2, but $\widehat{\mathsf{ISE}}_{LOO}(\eta_n)$ overestimates it by a

factor of 10. For the empirical mean $\overline{\eta_n} = \mathbf{1}_n^\top \mathbf{y}_n/n$, we obtain $\widehat{\mathsf{ISE}}_{LOO}(\overline{\eta_n}) \simeq 0.0758$, suggesting a better prediction of f by $\overline{\eta_n}$ than by η_n, whereas $\widehat{\mathsf{ISE}}_{BLP}(\overline{\eta_n}) \simeq 0.0610 > \widehat{\mathsf{ISE}}_{BLP}(\eta_n) \simeq 0.0053$, indicating that η_n is a better predictor than $\overline{\eta_n}$ (and indeed, $\mathsf{ISE}(\overline{\eta_n}) \simeq 0.0625 > \mathsf{ISE}(\eta_n) \simeq 0.0097$).

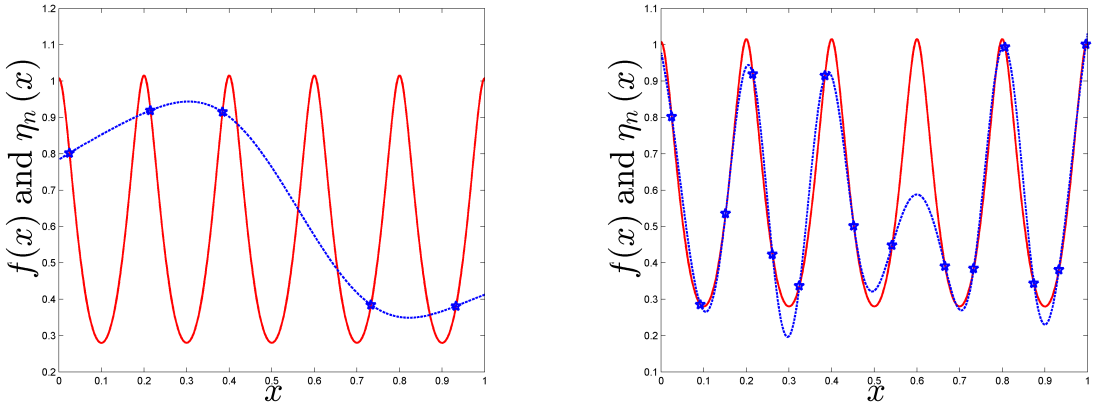

Figure 21: f and its interpolator η_n for $n = 5$ (left) and $n = 15$ (right).

H Matlab code for calculating $\widehat{\mathsf{ISE}}_{BLP}(\eta_n)$ and $\widehat{\mathsf{ISE}}_{BLUP}(\eta_n)$

```
function [ ise_BLP,ise_BLP_unbiased,ise_LOOCV,BLUE ] = ...
    ISE_WLOO_BLP( yn,Xn,Rn,Xtest,Mu,WnN,Kn_BLP,knx_BLP,nugget,constant_term )
% function [ ise_BLP,ise_BLP_unbiased,ise_LOOCV,BLUE ] = ...
%    ISE_WLOO_BLP( yn,Xn,Rn,Xtest,Mu,WnN,Kn_BLP,knx_BLP,nugget,constant_term )
% Xn = d*n matrix of design points
% yn = n (column) vector of observations
% Rn = n*n matrix such that LOO errors = Rn'*yn for the predictor evaluated
% Xtest = d*N matrix of test points (the ISE is estimated by the empirical
%   mean on Xtest)
% Mu = 1*N row vector of weights (with sum = 1) defining the measure on Xtest
% WnN = n*N matrix, whose ith column is the vector of weights at the ith
%   test point for the predictor evaluated (predictions over Xtest are
%   given by yn'*WnN)
% Kn_BLP = n*n kernel matrix, with the kernel used for ise_BLP estimation
% knx_BLP = n*N kernel matrix for the n design points and N test points
% nugget = nugget parameter (presence of an additive noise with variance
%   nugget*s2, with s2 the GP variance)
% ise_BLP = ISE estimated by construction of the BLP
% ise_BLP_unbiased = unbiased version of the above
% ise_LOOCV = classical LOOCV estimate (sum of squares of LOO errors)/n
% if constant_term == 1, the construction assumes that there is a constant
```

```
%   term in the model and estimates it to correct the ISE estimation
% BLUE = estimator of the constant term in the location model

[~,n]=size(Xn);
BLUE=NaN;
LOO_errors=Rn'*yn;
LOO_errors_squared=LOO_errors.^2;
ise_LOOCV=mean(LOO_errors_squared); % standard LOO ISE estimator
% squared error for the predictor and assumed model (as if no trend)
rhon2_12=1+nugget-2*sum(WnN.*knx_BLP,1)+sum(WnN.*(Kn_BLP*WnN),1);
tn_12=knx_BLP-Kn_BLP*WnN;
un=ones(n,1); Knm1un=Kn_BLP\un;
if constant_term == 1
    % include a constant term in the model for the BLP estimator
    constant=yn'*Knm1un/(un'*Knm1un); % = BLUE of the constant
    BLUE=constant;
    % remove the constant, shift the observations, and proceed as for a
    % model with zero mean, add a suitable constant to the estimated ISE
    yn=yn-constant*un;
    LOO_errors=Rn'*yn;
    LOO_errors_squared=LOO_errors.^2;
    ISE_add_constant=constant^2*mean((un'*WnN-1).^2);
end
Qdum=Rn'*Kn_BLP*Rn;
um=diag(Qdum);
cm=um*rhon2_12+2*(Rn'*tn_12).^2;
Sm=um*um'+2*Qdum.^2;
% weights for the BLP
wLOO=(Sm\cm);
% weights for the BLUP
wLOO_unbiased=Sm\(cm+um*(rhon2_12-um'*wLOO)/(um'*(Sm\um)));
error2_WLOOCV=LOO_errors_squared'*wLOO; error2_WLOOCV=max(error2_WLOOCV,0);
error2_WLOOCV_unbiased=LOO_errors_squared'*wLOO_unbiased;
        error2_WLOOCV_unbiased=max(error2_WLOOCV_unbiased,0);
ise_BLP=sum(Mu.*error2_WLOOCV);
ise_BLP_unbiased=sum(Mu.*error2_WLOOCV_unbiased);
if constant_term==1
    ise_BLP=ise_BLP+ISE_add_constant;
    ise_BLP_unbiased=ise_BLP_unbiased+ISE_add_constant;
end
end
```

References

[1] J. Loeppky, J. Sacks, and W. Welch, *Choosing the sample size of a computer ex-*

periment: a practical guide, Journal of the American Statistical Association, 51 (2009), pp. 366–376.

[2] L. Pronzato, *Sensitivity analysis via Karhunen-Loève expansion of a random field model: estimation of Sobol' indices and experimental design*, Reliability Engineering and System Safety, 187 (2019), pp. 93–109.

[3] L. Pronzato and M.-J. Rendas, *Bayesian local kriging*, Technometrics, 59 (2017), pp. 293–304.

[4] S. Sarykalin, G. Serraino, and S. Uryasev, *Value-at-risk vs. conditional value-at-risk in risk management and optimization*, in State-of-the-art decision-making tools in the information-intensive age, Informs, 2008, pp. 270–294.