Model parameters in Gaussian process interpolation: an empirical study of selection criteria

Sébastien Petit, Julien Bect, Paul Feliot, Emmanuel Vazquez

Model parameters in Gaussian process interpolation: an empirical study of selection criteria

Sébastien J. Petit[†*], Julien Bect[†], Paul Feliot[*‡], and Emmanuel Vazquez[†]

Abstract. This article revisits the fundamental problem of parameter selection for Gaussian process interpolation. By choosing the mean and the covariance functions of a Gaussian process within parametric families, the user obtains a family of Bayesian procedures to perform predictions about the unknown function, and must choose a member of the family that will hopefully provide good predictive performances. We base our study on the general concept of scoring rules, which provides an effective framework for building leave-one-out selection and validation criteria, and a notion of extended likelihood criteria based on an idea proposed by Fasshauer and co-authors in 2009, which makes it possible to recover standard selection criteria such as, for instance, the generalized cross-validation criterion. Under this setting, we empirically show on several test problems of the literature that the choice of an appropriate family of models is often more important than the choice of a particular selection criterion (e.g., the likelihood versus a leave-one-out selection criterion). Moreover, our numerical results show that the regularity parameter of a Matérn covariance can be selected effectively by most selection criteria.

Key words. Gaussian processes interpolation, Model choice; Parameter selection; Scoring rules; Likelihood; Regularity

AMS subject classifications. 62G08, 65D05

1. Introduction.

Regression and interpolation with Gaussian processes, or kriging, is a popular statistical tool for non-parametric function estimation, originating from geostatistics and time series analysis, and later adopted in many other areas such as machine learning and the design and analysis of computer experiments (see, e.g., [39, 40, 42] and references therein). It is widely used for constructing fast approximations of time-consuming computer models, with applications to calibration and validation [5, 30], engineering design [21, 29], Bayesian inference [9, 48], and the optimization of machine learning algorithms [7]—to name but a few.

A Gaussian process (GP) prior is characterized by its mean and covariance functions. They are usually chosen within parametric families (for instance, constant or linear mean functions, and Matérn covariance functions), which transfers the problem of choosing the mean and covariance functions to that of selecting parameters. The selection is most often carried out by optimization of a criterion that measures the goodness of fit of the predictive distributions, and a variety of such criteria—the likelihood function, the leave-one-out (LOO) squared-prediction-error criterion (hereafter denoted by LOO-SPE), and others—is available from the literature. The search for arguments to guide practitioners in the choice of one particular criterion is the main motivation of this study.

As a necessary parenthesis, note that the fully Bayesian statistician does not select a particular

[*]Safran Aircraft Engines, Moissy-Cramayel, France (sebastien.petit@safrangroup.com). This work of S. Petit was funded by the french *Agence Nationale de la Recherche et de la Technologie* (ANRT) under a CIFRE grant.

[†]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France (sebastien.petit@centralesupelec.fr, julien.bect@centralesupelec.fr, emmanuel.vazquez@centralesupelec.fr).

[‡]P. Feliot no longer works at Safran Aircraft Engines. He can be contacted at p.feliot@hotmail.fr.

value for the parameters, but chooses instead to average the predictions over a *posterior* distribution on the parameters [27]. This approach may be preferred for more robust predictions (see, e.g., [6]), but comes at a higher computational cost. For this reason, the present article will set aside the fully Bayesian approach and concentrate on the plugin approach, where only one parameter value is chosen to carry out predictions.

The two most popular methods for parameter selection are maximum likelihood (ML) and cross-validation (CV) based on LOO criteria, which were introduced in the field of computer experiments by the seminal work of [14]. Since then, despite a fairly large number of publications dealing with ML and CV techniques for the selection of a GP model, the literature has remained in our view quite sparse about the relative merits of these methods, from both theoretical and empirical perspectives.

For instance, in the framework of interpolation and infill asymptotics, where observations accumulate in some bounded domain, [50] and [51] show that some combinations of the parameters of a Matérn covariance function can be estimated consistently by ML. Again in the case of infill asymptotics, with the additional assumption of a one-dimensional GP with an exponential covariance function, [4] show that estimation by LOO is also consistent. (Similar results exist in the expanding asymptotic framework, where observations extend over an ever-increasing horizon.)

The practical consequences of the aforementioned results are somewhat limited in our view because practitioners are primarily interested in the quality of the predictions. Knowing that the parameters of a GP can be estimated consistently is intellectually reassuring, but may be considered of secondary importance. These results are indeed about the statistical model itself but they say little about the prediction properties of the model. Besides, there does not exist at present, to our knowledge, some theoretically-based evaluation of the relative performances of ML and CV selection criteria under infill asymptotics.

Turning now to empirical comparisons of selection criteria for GP interpolation, the first attempt in the domain of computer experiments can be traced back to the work of [14, 15]. The authors introduce CV and ML—which can be seen as a special kind of cross-validation—and present some simple experiments using tensorized covariance functions, from which they conclude that, “Of the various kinds of cross-validation [they] have tried, maximum likelihood seems the most reliable”. Additional experiments have been conducted by several authors, but no consensus emerges from these studies: [2, 31, 43] conclude in favor of CV whereas [32] advocate ML.

These studies are limited to a rather small number of test functions and covariance functions, which may explain the discrepancy in the conclusions of those experiments. In particular, only [2] considers the popular and versatile Matérn covariance functions. Moreover, most studies focus only on the accuracy of the posterior mean—only [43] and [2] provide results accounting for the quality of the posterior variance—whereas the full posterior predictive distribution is used in most GP-based methods (see, e.g., [10, 29]).

This article presents two main contributions. First, we improve upon the results of the literature by providing an empirical ranking of selection criteria for GP interpolation, according to several metrics measuring the quality of posterior predictive distributions on a large set of test functions from the domain of computer experiments. To this end, we base our study on the general concept of scoring

rules [24, 52], which provides an effective framework for building selection and validation criteria. We also introduce a notion of extended likelihood criteria, borrowing an idea from Fasshauer and co-authors [19, 20] in the literature of radial basis functions.

Second, we provide empirical evidence that the choice of an appropriate family of models is often more important—and sometimes much more important, especially when the size of the design increases—than the choice of a particular selection criterion (e.g., likelihood versus LOO-SPE). More specifically, in the case of the Matérn family, this leads us to assess, and ultimately recommend, the automatic selection of a suitable value of the regularity parameter, against the common practice of choosing beforehand an arbitrary value of this parameter.

The article is organized as follows. Section 2 briefly recalls the general framework of GP regression and interpolation. Section 3 reviews selection criteria for GP model parameters. After recalling the general notion of scoring rules, we present a broad variety of selection criteria from the literature. Section 4 presents experimental results on the relative performances of these criteria. Section 5 presents our conclusions and perspectives.

2. General framework.

Let us consider the general GP approach for a scalar-valued deterministic computer code with input space $\mathbb{X} \subseteq \mathbb{R}^d$. The output of the computer code $z : \mathbb{X} \to \mathbb{R}$ is modeled by a random function $(Z(x))_{x\in\mathbb{X}}$, which, from a Bayesian perspective, represents prior knowledge about z. If we assume that $Z(\cdot)$ is observed on a design $\mathbb{X}_n = \{x_1, \ldots, x_n\}$ of size n, the data corresponds to a sample of the random vector $\mathbf{Z}_n = (Z(x_1), \ldots, Z(x_n))^\mathsf{T}$.

The GP assumption makes it possible to easily derive posterior distributions. More precisely, it is assumed that $Z(\cdot)$ is a Gaussian process, with (prior) mean $\mathsf{E}(Z(x)) = \sum_{l=1}^{L} \beta_l \phi_l(x)$, where the $\beta_1, \ldots, \beta_L$ are unknown regression parameters and $\phi_1, \ldots, \phi_l$ are known regression functions, and with (prior) covariance $\mathrm{cov}\,(Z(x), Z(y)) = k_\theta(x, y)$, where $\theta \in \Theta \subseteq \mathbb{R}^q$ is a vector of *parameters*. Throughout the article, the covariance matrix of $\mathbf{Z}_n$ will be denoted by $\mathbf{K}_\theta$. We assume for simplicity that the prior mean of $Z(\cdot)$ is zero (hence, $L = 0$), which is a common practice when data are centered.

One of the most popular covariance functions for GP regression is the anisotropic stationary Matérn covariance function [33] popularized by [42]:

$$k_\theta(x, y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} h\right)^\nu \mathscr{K}_\nu \left(\sqrt{2\nu} h\right), \quad h = \left(\sum_{j=1}^{d} \frac{(x_j - y_j)^2}{\rho_j^2} \right)^{1/2}, \tag{2.1}$$

where Γ is the Gamma function, $\mathscr{K}_\nu$ is the modified Bessel function of the second kind, and θ denotes the vector of parameters $\theta = (\sigma^2, \rho_1, \ldots, \rho_d, \nu) \in \Theta =]0, \infty[^{d+2}$. The parameter σ^2 is the variance of $Z(\cdot)$, the ρ_is are range parameters which characterize the typical correlation length on each dimension, and ν is a regularity parameter, whose value controls the mean-square differentiability of $Z(\cdot)$. Recall (see Table 1) that the Matérn covariance function with $\nu = 1/2$ corresponds to the so-called exponential covariance function, and the limiting case $\nu \to \infty$ can be seen as the "squared exponential" (also called Gaussian) covariance function.

Because $\mathscr{K}_\nu$ has a closed-form expression when $\nu - \frac{1}{2}$ is an integer, and is more expensive to evaluate numerically in other cases, most implementations choose to restrict ν to half-integer values. Moreover, a widespread practice (in applications and research papers) consists in selecting a particular

Table 1: Popular Matérn subfamilies

	$\nu = \frac{1}{2}$	$\nu = \frac{3}{2}$	$\nu = \frac{5}{2}$	$\nu = +\infty$
a.k.a.	exponential			squared exponential
$k_\theta(x,y)$	$\sigma^2 e^{-h}$	$\sigma^2(1+\sqrt{3}h)e^{-\sqrt{3}h}$	$\sigma^2(1+\sqrt{5}h+\frac{5h^2}{3})e^{-\sqrt{5}h}$	$\sigma^2 e^{-\frac{h^2}{2}}$

value for ν (e.g., $\nu = 1/2$, $\nu = 3/2\ldots$ or the limiting case $\nu \to \infty$), once and for all.

Since the family of Matérn covariance functions is widely used in practice, we focus exclusively on this family in this work. We believe that the conclusions of the present study would not be altered significantly if other families of covariance functions (e.g., the compactly supported covariance functions proposed by [47]) were considered.

Once a GP model has been chosen, the framework of Gaussian process regression allows one to build a predictive distribution $\mathcal{N}(\mu_\theta(x), \sigma^2_\theta(x))$ for an unobserved $Z(x)$ at $x \in \mathbb{R}^d$, where

$$\left\{ \begin{array}{rcl} \mu_\theta(x) & = & \mathbf{k}^*_\theta(x)^\mathsf{T} \mathbf{K}_\theta^{-1} \mathbf{Z}_n, \\ \sigma^2_\theta(x) & = & k_\theta(x,x) - \mathbf{k}^*_\theta(x)^\mathsf{T} \mathbf{K}_\theta^{-1} \mathbf{k}^*_\theta(x) \end{array} \right. \tag{2.2}$$

with $\mathbf{k}^*_\theta(x) = (k_\theta(x,x_1), \ldots, k_\theta(x,x_n))^\mathsf{T}$. More generally, predictive distributions can be built for a larger range of quantities of interest such as joint observations, derivatives, integrals or excursions of Z above a given threshold.

Using this framework, the user obtains a family of Bayesian procedures, indexed by θ, to perform predictions about the unknown computer code at hand, and must choose a member of the family that will hopefully provide good predictive performances.

3. Selection of a GP model from a parameterized family.

3.1. Scoring rules.

Goodness-of-fit criteria for probabilistic predictions have been studied in the literature under the name of scoring rules by [24]. A (negatively oriented) scoring rule is a function $S(\cdot\,; z) : \mathscr{P} \to \mathbb{R} \cup \{-\infty, +\infty\}$, acting on a class $\mathscr{P}$ of probability distributions on $\mathbb{R}$, such that $S(P; z)$ assigns a loss for choosing a predictive distribution $P \in \mathscr{P}$, while observing $z \in \mathbb{R}$. Scoring rules make it possible to quantify the quality of probabilistic predictions.

Example 1 (**squared prediction error**). *Denoting by μ the mean of a predictive distribution P, the squared prediction error*

$$S^{\mathrm{SPE}}(P; z) = (z - \mu)^2 \tag{3.1}$$

accounts for the deviation of z from μ. Note that S^{SPE} ignores subsequent moments and therefore predictive uncertainties.

Example 2 (**negative log predictive density**). *Denoting by p the probability density of P,*

$$S^{\mathrm{NLPD}}(P; z) = -\log(p(z)) \tag{3.2}$$

Table 2: Scoring rules behavior as $|\mu - z| \ll 1$.

	$\sigma \ll \lvert\mu - z\rvert$	$\sigma \simeq \lvert\mu - z\rvert$	$\sigma \gg \lvert\mu - z\rvert$
$S^{\mathrm{SPE}}(P; z)$	0	0	0
$S^{\mathrm{NLPD}}(P; z)$	$+\infty$	$-\infty$	$\log(2\pi\sigma)$
$S^{\mathrm{CRPS}}(P; z)$	0	0	$\propto \sigma$
$S^{\mathrm{IS}}_{1-\alpha}(P; z)$	0	0	$\propto \sigma$

tells how likely z is according to P. [8] shows that any (proper) scoring rule that depends on $p(z)$ (and only on $p(z)$) can be reduced to S^{NLPD}.

Example 3 (**continuous ranked probability score**). *Let U and U' be two independent random variables with distribution P. The CRPS quantifies the deviation of U from z:*

$$S^{\mathrm{CRPS}}(P; z) = \mathsf{E}(|U - z|) - \frac{1}{2}\mathsf{E}(|U - U'|)\,. \tag{3.3}$$

Since it is possible to show that $S^{\mathrm{CRPS}}(P; z) = \int (P(U \leq u) - \mathbb{1}_{z \leq u})^2 \,\mathrm{d}u$, the CRPS can also be seen as a (squared) distance between the empirical cumulative distribution $u \mapsto \mathbb{1}_{z \leq u}$ and the cumulative distribution of P.

Note that if absolute values in (3.3) are replaced by squared values, then S^{SPE} is recovered. The CRPS can also be extended to the so-called energy and kernel scores [24] by observing that $(x, y) \mapsto |x - y|$ is a conditionally negative kernel.

Example 4 (**interval score**). *The interval scoring rule at level $1 - \alpha$ is defined, for $\alpha \in]0, 1[$, by*

$$S^{\mathrm{IS}}_{1-\alpha}(P; z) = (u - l) + \frac{2}{\alpha}(l - z)\mathbb{1}_{z \leq l} + \frac{2}{\alpha}(z - u)\mathbb{1}_{z > u} \tag{3.4}$$

where l and u are the $\alpha/2$ and $1 - \alpha/2$ quantiles of P. The first term penalizes large intervals, while the second and third terms penalize intervals not containing z.

When the predictive distribution P is Gaussian, which is the case when P is the posterior distribution of a GP Z at a given point, the aforementioned scoring rules all have closed-form expressions. More precisely, for $P = \mathcal{N}(\mu, \sigma^2)$, we simply have $S^{\mathrm{SPE}}(P; z) = (z - \mu)^2$ and $S^{\mathrm{NLPD}}(P; z) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}(z - \mu)^2/\sigma^2$. $S^{\mathrm{IS}}_{1-\alpha}$ can be obtained by taking the standard expressions of the $\alpha/2$ and $1 - \alpha/2$ quantiles of P, and it can be shown that

$$S^{\mathrm{CRPS}}(P; z) = \sigma\Big(\frac{z - \mu}{\sigma}\Big(2\Phi\big(\frac{z - \mu}{\sigma}\big) - 1\Big) + 2\phi\big(\frac{z - \mu}{\sigma}\big) - \frac{1}{\sqrt{\pi}}\Big),$$

where ϕ and Φ stand respectively for the probability density function and the cumulative distribution function of the standard Gaussian distribution.

Note that all aforementioned scoring rules penalize large values of $|z - \mu|$. When $|z - \mu| \ll 1$ different scoring rules yield different penalties, as reported in Table 2.

3.2. Selection criteria.

3.2.1. Leave-one-out selection criteria. Scoring rules make it possible to build criteria for choosing the parameters of a GP. More precisely, to select θ based on a sample $Z_1, \ldots, Z_n$, one can minimize the mean score

$$J_n^S(\theta) = \frac{1}{n}\sum_{i=1}^{n} S(P_{\theta,-i}; Z_i), \tag{3.5}$$

where S is a scoring rule and $P_{\theta,-i}$ denotes the distribution of Z_i conditional on the Z_js, for $1 \leq j \leq n,\ j \neq i$, indexed by θ.

Selection criteria written as (3.5) correspond to the well-established leave-one-out (LOO) method, which has been proposed in the domain of computer experiments by [14], and is now used in many publications (see, e.g., [39], and also [52], who formally adopt the point of view of the scoring rules, but for model validation instead of parameter selection).

Efficient computation of predictive distributions. Leave-one-out predictive densities can be computed using fast algebraic formulas [11, 18]. More precisely, the predictive distribution $P_{\theta,-i}$ is a normal distribution $\mathcal{N}(\mu_{\theta,-i}, \sigma^2_{\theta,-i})$ with

$$\mu_{\theta,-i} = Z(x_i) - \frac{(\mathbf{K}_\theta^{-1}\mathbf{Z}_n)_i}{\mathbf{K}^{-1}_{\theta,i,i}} \quad \text{and} \quad \sigma^2_{\theta,-i} = \frac{1}{\mathbf{K}^{-1}_{\theta,i,i}}. \tag{3.6}$$

Furthermore, [36] show that, using reverse-mode differentiation, it is possible to compute mean scores J_n and their gradients with a $O(n^3 + dn^2)$ computational cost, which is the same computational complexity as for computing the likelihood function and its gradient (see, e.g., [39]).

The particular case of LOO-SPE. The LOO selection criterion

$$J_n^{\mathrm{SPE}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} (\mu_{\theta,-i} - Z(x_i))^2, \tag{3.7}$$

based on the scoring rule (3.1) will be referred to as LOO-SPE. This criterion, also called prediction sum of squares (PRESS) [1, 45] or LOO squared bias [14], is well known in statistics and machine learning, and has been advocated by some authors [2,3,31,40,43] to address the case of "misspecified" covariance functions.

However, note that σ^2 cannot be selected using J_n^{SPE}. When J_n^{SPE} is used, σ^2 is generally chosen (see, e.g., [2, 14] and Remark 3.1) to satisfy

$$\frac{1}{n}\sum_{i=1}^{n} \frac{(Z(x_i) - \mu_{\theta,-i})^2}{\sigma^2_{\theta,-i}} = 1, \tag{3.8}$$

which will be referred to as Cressie's rule for σ^2, in reference to the claim by [12] that (3.8) should hold approximately for a good GP model.

Other scoring rules for LOO. The selection criteria using the NLPD scoring rule (3.2) and the CRPS scoring rule (3.3) will be referred to as the LOO-NLPD and LOO-CRPS criteria, respectively. The LOO-NLPD criterion has been called preditive deficiency in [14], and Geisser's surrogate Predictive Probability (GPP) in [43]. The LOO-CRPS criterion has been considered in [52] as a criterion for model validation (see also [16] for an application to model selection), and more recently [36, 37] as a possible criterion for parameter selection as well.

Remark 3.1. Note that Cressie's rule (3.8) can be derived by minimizing the LOO-NLPD criterion with respect to σ^2.

Remark 3.2. In order to limit the number of selection criteria under study, the interval scoring rule is only used for validation in this work.

3.2.2. Maximum likelihood and generalizations. We can safely say that the most popular method for selecting θ from data is maximum likelihood estimation—and related techniques, such as restricted maximum likelihood estimation. The ML estimator is obtained by maximizing the likelihood function or, equivalently, by minimizing the negative log-likelihood (NLL) selection criterion. Denoting by $p_\theta(\mathbf{Z}_n)$ the joint density of $\mathbf{Z}_n$, the NLL selection criterion may be written as

$$J_n^{\mathrm{NLL}}(\theta) = -\log(p_\theta(\mathbf{Z}_n)) = \frac{1}{2}\left(n\log(2\pi) + \log\det\mathbf{K}_\theta + \mathbf{Z}_n^\top \mathbf{K}_\theta^{-1}\mathbf{Z}_n\right). \tag{3.9}$$

As pointed out by [14], the NLL criterion is closely related to the LOO-NLPD criterion, through the identity

$$J_n^{\mathrm{NLL}}(\theta) = -\log(p_\theta(Z(x_1)) - \sum_{i=2}^{n} \log(p_\theta(Z(x_i) \mid Z(x_1),\ldots,Z(x_{i-1}))),$$

where the predictive distributions of the $Z(x_i)$s given the $Z(x_1),\ldots,Z(x_{i-1})$ explicitly appear.

One can minimize (3.9) in closed-form with respect to σ^2, given other parameters. Writing $\mathbf{K}_\theta = \sigma^2\mathbf{R}_\theta$ and canceling $\partial J_n^{\mathrm{NLL}}(\theta)/\partial\sigma^2 = \left(n\sigma^2 - \mathbf{Z}_n^\top\mathbf{R}_\theta^{-1}\mathbf{Z}_n\right)/(2\sigma^2)$ yields

$$\sigma^2_{\mathrm{NLL}} = \frac{1}{n}\mathbf{Z}_n^\top\mathbf{R}_\theta^{-1}\mathbf{Z}_n, \tag{3.10}$$

which will be referred to as the profiling rule for σ^2.

Injecting (3.10) into (3.9) yields a *profiled likelihood* selection criterion, that can be written as

$$J_n^{\mathrm{PL}}(\theta) = \log\sigma^2_{\mathrm{NLL}} + \frac{1}{n}\log\det\mathbf{R}_\theta = \log\left(\frac{1}{n}\mathbf{Z}_n^\top\mathbf{R}_\theta^{-1}\mathbf{Z}_n\right) + \frac{1}{n}\log\det\mathbf{R}_\theta\,. \tag{3.11}$$

Following Fasshauer and co-authors [19, 20], we consider now a family of selection criteria that extends (3.9). Using the factorization $\mathbf{R}_\theta = \mathbf{Q}\Lambda\mathbf{Q}^\top$, where $\mathbf{Q} = (\mathbf{q}_1,\ldots,\mathbf{q}_n)$ is an orthogonal matrix of (orthonormal) eigenvectors and $\Lambda = \mathrm{diag}(\lambda_1,\cdots,\lambda_n)$, notice that

$$\exp\left(J_n^{\mathrm{PL}}(\theta)\right) = \frac{1}{n}\mathbf{Z}_n^\top\mathbf{R}_\theta^{-1}\mathbf{Z}_n\cdot(\det\mathbf{R}_\theta)^{1/n} \propto \left(\sum_{i=1}^{n}\left(\mathbf{q}_i^\top\mathbf{Z}_n\right)^2/\lambda_i\right)\left(\prod_{i=1}^{n}\lambda_j\right)^{1/n}. \tag{3.12}$$

This suggests a generalization of the likelihood criterion that we shall call Fasshauer's Hölderized likelihood (HL), defined as

$$J_n^{\mathrm{HL},p,q}(\theta) = \left(\sum_{i=1}^{n}(\mathbf{q}_i^\top\mathbf{Z}_n)^2/\lambda_i^p\right)^{1/p}\left(\frac{1}{n}\sum_{j=1}^{n}\lambda_j^q\right)^{1/q}, \tag{3.13}$$

with $q\in[-\infty,+\infty]$, and $p\in\mathbb{R}\setminus\{0\}$, and where σ^2 can be chosen using the rules (3.8) or (3.10), since $J_n^{\mathrm{HL},p,q}(\theta)$ does not depend on σ^2. Owing to the standard property of generalized means $\left(\frac{1}{n}\sum_{i=1}^n x_i^q\right)^{\frac{1}{q}} \overset{q\to 0}{\longrightarrow} \sqrt[n]{x_1\cdots x_n}$), (3.12) is recovered by taking $p=1$ and letting $q\to 0$. Moreover, two other known selection criteria can be obtained for particular values of p and q, as detailed below.

Generalized cross-validation. Taking $p = 2$ and $q = -1$ in (3.13) yields the generalized cross-validation (GCV) criterion

$$J_n^{\mathrm{GCV}}(\theta) = n^{-1} \left(J_n^{\mathrm{HL},2,-1}(\theta) \right)^2,$$

which was originally proposed as a rotation-invariant version of PRESS [25] for linear models. It has also been shown to be efficient for the selection of the smoothing parameter of polyharmonic splines [45] and for the selection of the degree of a spline [46].

The GCV selection criterion is a weighted SPE criterion, which can also be written as

$$J_n^{\mathrm{GCV}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} w_i^2(\theta) (Z(x_i) - \mu_{\theta,-i})^2, \quad w_i(\theta) = \frac{\tilde{\sigma}^2}{\sigma_{\theta,-i}^2}, \tag{3.14}$$

with $\tilde{\sigma}^2 = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma_{\theta,-i}^2} \right)^{-1}$. Notice that $w_i(\theta)$ is lower when $\sigma_{\theta,-i}$ is larger, which happens when points are either isolated or lying on the border / envelope of $\mathbb{X}_n$. Equation (3.14) shows that, similarly to the LOO criteria of Section 3.2.1, the GCV criterion can be computed, along with its gradient, in $O(n^3 + dn^2)$ time. However, since $\tilde{\sigma}$ depends on the data, the J_n^{GCV} criterion cannot be formally derived from a scoring rule.

Kernel alignment. The kernel alignment selection criterion is defined as

$$J_n^{\mathrm{KA}}(\theta) = -\frac{\mathbf{Z}_n^{\mathsf{T}} \mathbf{K}_\theta \mathbf{Z}_n}{\|\mathbf{K}_\theta\|_F \|\mathbf{Z}_n\|^2}, \tag{3.15}$$

where $\|\cdot\|_F$ stands for the Frobenius matrix norm. This criterion can be derived from (3.13) by taking $p = -1$ and $q = 2$:

$$J_n^{\mathrm{KA}}(\theta) = -\frac{1}{\sqrt{n} \|\mathbf{Z}_n\|^2 J_n^{\mathrm{HL},-1,2}(\theta)}.$$

It was originally proposed in the machine learning literature [13] to maximize the alignment of $\mathbf{Z}_n$ with the eigenvector of $\mathbf{K}_\theta$ corresponding to the largest eigenvalue. This criteria is noticeably cheaper than the others, as it does not require to invert $\mathbf{K}_\theta$ and can therefore be computed along with its gradient in $O(dn^2)$ time.

Remark 3.3. We choose to focus in this article on three well-known selection criteria (NLL, GCV and KA) that can be seen as special cases of (3.13), corresponding repectively to (p,q) equal to $(1,0)$, $(2,-1)$ and $(-1,2)$. The study of new selection criteria, obtained for other values of (p,q), is left for future work.

3.3. Hybrid selection criteria.

When considering several parameterized models—or, equivalently, when dealing with discrete parameters, such as half-integer values for the regularity parameter of the Matérn covariance—some authors suggest to use one selection criterion to select the parameters in each particular model, and a different one to select the model itself.

For instance, in [29], the authors select the parameters of a power-exponential covariance function using the NLL selection criterion (i.e., the ML method), and then select a suitable transformation of the output of the simulator, in a finite list of possible choices, using the LOO-SPE criterion. Similarly,

the NLL selection criterion is combined in [16] with a variety of model-validation criteria, including LOO-CRPS and LOO-NLPD.

In Section 4 we will denote by NLL/SPE the hybrid method that selects the variance and range parameters of a Matérn covariance function using the NLL criterion, and then minimizes the LOO-SPE criterion to select the regularity parameter ν in finite list of values.

4. Numerical experiments.

4.1. Methodology.

We investigate the problem of parameter selection with an experimental approach consisting of four ingredients: 1) a set of unknown functions f to be predicted using evaluation results on a finite design $\mathbb{X}_n = \{x_1, \ldots, x_n\} \subset \mathbb{X}$; 2) the GP regression method that constructs predictive distributions $P_{\theta,x}$ of f at given xs in $\mathbb{X}$, indexed by parameter θ; 3) several selection criteria J_n to choose θ; 4) several criteria to assess the quality of the $P_{\theta,x}$s. Details about each of these ingredients are given below (starting from the last one).

Criteria to assess the quality of the $P_{\theta,x}$s. A natural way to construct a criterion to assess the quality of the $P_{\theta,x}$s is to choose a scoring rule S and to consider the mean score on a test set $\mathbb{X}_N^{\text{test}} = \{x_1^{\text{test}}, \ldots, x_N^{\text{test}}\} \subset \mathbb{X}$ of size N:

$$R(\theta; S) = \frac{1}{N} \sum_{i=1}^{N} S\big(P_{\theta, x_i^{\text{test}}}; f(x_i^{\text{test}})\big)\,. \tag{4.1}$$

Selection criteria. We shall consider the selection criteria J_n presented in Section 3, namely, the LOO-SPE, LOO-NLPD, LOO-CRPS, NLL, GCV, KA and NLL/SPE selection criteria. Given a function f and a design $\mathbb{X}_n$, each selection criterion J_n yields a parameter θ_{J_n}.

Parameterized GP models. In this work, models are implemented using a custom version of the [26] Python package (see Supplementary Material, hereafter abbreviated as SM). We assume no observation noise, which corresponds to the interpolation setting. All functions will be centered beforehand to have zero-mean on $\mathbb{X}_N^{\text{test}}$, and we will consider zero-mean GPs only. The anisotropic Matérn covariance function (2.1) is used, with parameter $\theta = (\sigma^2, \rho_1, \ldots, \rho_d, \nu)$, and the regularity parameter ν is either set a priori to $\nu = \chi + 1/2$, with $\chi \in \{0, 1, 2, 3, 4, d, 2d, \infty\}$, or selected automatically. The latter case will be denoted by $\widehat{\nu}$.

Remark 4.1. Since the covariance matrix of $\mathbf{Z}_n$ can be nearly singular when the range parameters take large values, we define upper bounds on these values in order to avoid the use of *nugget* or *jitter* (see, e.g., [35, 38]). Details are provided in the SM.

Test functions. The test functions used in the study are described in the next section. They are grouped into collections, and we provide averaged values of mean-score metrics of the form (4.1) for each collection.

4.2. Test functions.

4.2.1. Design of a low-pass filter.

Fuhrländer and Schöps [22] consider the problem of computing, using a frequency-domain PDE solver, the scattering parameters S_ω of an electronic component called stripline low-pass filter, at several values of the excitation pulsation ω. The geometry of the

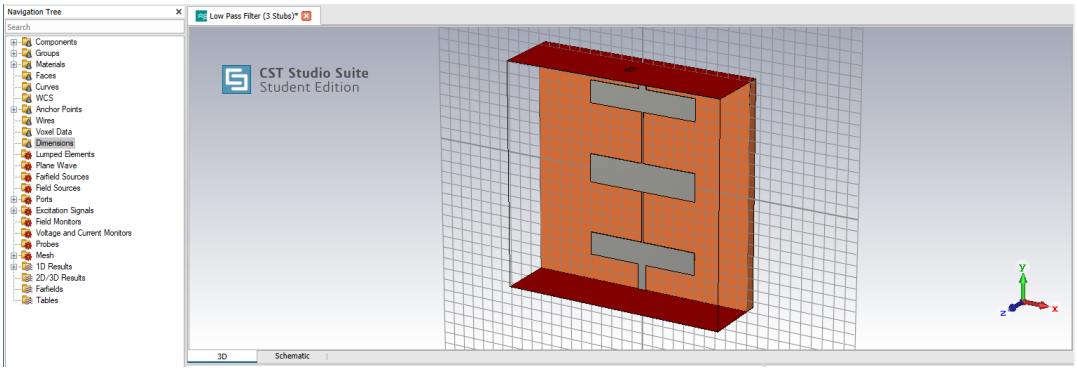

Figure 1: A low-pass filter design problem in CST Studio Suite®.

stripline filter is illustrated on Figure 1. It is parameterized using six real valued factors concatenated in a vector $x \in \mathbb{R}^d$, $d = 6$. The objective is to satisfy the low-pass specifications $|S_{2k\pi}(x)| \geq -1\text{dB}$ for $0 \leq k \leq 4$ and $|S_{2k\pi}(x)| \leq -20\text{dB}$ for $5 \leq k \leq 7$. Meeting such requirements is a difficult and time-consuming task.

In this article we consider the quantities $\text{Re}(S_{2\pi})$, $\text{Re}(S_{6\pi})$, $\text{Re}(S_{10\pi})$ and $\text{Re}(S_{14\pi})$. We randomly sample $M = 100$ designs $\mathbb{X}_n$ of size $n = 300$ from a database of 10000 simulation results, and use the remaining $N = 10000 - n = 9700$ points as test sets. The metric (4.1) is computed and averaged on these M test sets.

4.2.2. Other test functions. We supplement the above engineering problem with a collection of test functions from the literature. More precisely, we consider the Goldstein-Price function [17], a one-dimensional version of the Goldstein-Price function (see SM for details), the Mystery function [31], the Borehole function [49], several collections obtained from the GKLS simulator [23], and the rotated Rosenbrock collection from the BBOB benchmark suite [28].

The GKLS simulator has a "smoothness" parameter $k \in \{0,1,2\}$ controlling the presence of non-differentiabilities on some nonlinear subspaces—the trajectories being otherwise infinitely differentiable. For both GKLS and Rosenbrock, two different values of the input dimension were considered ($d = 2$ and $d = 5$). The resulting set of twelve problems—considering that changing the value of k or d defines a new problem—is summarized in Table 3.

For each problem, we consider three design sizes $n \in \{10d, 20d, 50d\}$. For the GKLS and Rosenbrock collections, we directly used the collections of test functions provided by the authors ($M = 100$ and 15 functions, respectively). For a given dimension, they are all evaluated on the same space-filling designs $\mathbb{X}_n$. For each of the remaining problems, we used a single test function, evaluated on $M = 100$ random space-filling designs $\mathbb{X}_n$, thereby creating collections of 100 data sets.

A Sobol' sequence $\mathbb{X}_N^{\text{test}}$ of size $N = 10000$ is used as test set and the functions are centered and normalized to unit variance on these test sets.

4.3. Results and findings.

Table 3: Twelve benchmark problems

Problem	Goldstein-Price	Mystery	$\text{GKLS}_{k=0}$	$\text{GKLS}_{k=1}$	$\text{GKLS}_{k=2}$	Rosenbrock	Borehole
d	$\{1,2\}$	2	$\{2,5\}$	$\{2,5\}$	$\{2,5\}$	$\{2,5\}$	8

Table 4: Averages (over the $M = 100$ designs) of the $R(\theta;S^{\text{SPE}})$ values for $\text{Re}(S_{6\pi})$ with $n = 10d = 60$. Optimal $R^\star$ values are given in the rightmost column for comparison. For comparison also, $R(\theta;S^{\text{SPE}}) = 3.26\cdot 10^{-4}$ for the $J_n^{\text{NLL/SPE}}$ selection criterion, which also selects ν (see Section 3.3). The gray scale highlights the order of magnitude of the discrepancies.

Scoring rule: S^{SPE}	NLL	LOO-SPE	LOO-NLPD	LOO-CRPS	KA	GCV	$R^\star$
$\nu = 1/2$	$4.94\cdot 10^{-2}$	$5.11\cdot 10^{-2}$	$4.84\cdot 10^{-2}$	$4.84\cdot 10^{-2}$	$4.29\cdot 10^{-1}$	$4.73\cdot 10^{-2}$	$4.44\cdot 10^{-2}$
$\nu = 3/2$	$3.85\cdot 10^{-3}$	$4.24\cdot 10^{-3}$	$3.52\cdot 10^{-3}$	$3.59\cdot 10^{-3}$	$3.82\cdot 10^{-1}$	$3.45\cdot 10^{-3}$	$2.97\cdot 10^{-3}$
$\nu = 5/2$	$4.02\cdot 10^{-4}$	$5.11\cdot 10^{-4}$	$4.18\cdot 10^{-4}$	$4.31\cdot 10^{-4}$	$3.71\cdot 10^{-1}$	$4.93\cdot 10^{-4}$	$3.21\cdot 10^{-4}$
$\nu = 7/2$	$2.88\cdot 10^{-4}$	$4.33\cdot 10^{-4}$	$3.73\cdot 10^{-4}$	$3.86\cdot 10^{-4}$	$3.54\cdot 10^{-1}$	$4.75\cdot 10^{-4}$	$2.26\cdot 10^{-4}$
$\nu = 9/2$	$2.96\cdot 10^{-4}$	$4.39\cdot 10^{-4}$	$4.22\cdot 10^{-4}$	$3.95\cdot 10^{-4}$	$3.42\cdot 10^{-1}$	$5.44\cdot 10^{-4}$	$2.26\cdot 10^{-4}$
$\nu = 13/2$	$3.15\cdot 10^{-4}$	$4.80\cdot 10^{-4}$	$4.48\cdot 10^{-4}$	$4.25\cdot 10^{-4}$	$3.29\cdot 10^{-1}$	$6.43\cdot 10^{-4}$	$2.32\cdot 10^{-4}$
$\nu = 25/2$	$3.46\cdot 10^{-4}$	$5.29\cdot 10^{-4}$	$4.92\cdot 10^{-4}$	$4.61\cdot 10^{-4}$	$3.14\cdot 10^{-1}$	$7.43\cdot 10^{-4}$	$2.45\cdot 10^{-4}$
$\nu = \infty$	$3.80\cdot 10^{-4}$	$6.34\cdot 10^{-4}$	$5.38\cdot 10^{-4}$	$5.38\cdot 10^{-4}$	$2.94\cdot 10^{-1}$	$8.75\cdot 10^{-4}$	$2.60\cdot 10^{-4}$
$\nu \in \{1/2,\cdots,\infty\}$	$3.07\cdot 10^{-4}$	$4.90\cdot 10^{-4}$	$4.64\cdot 10^{-4}$	$4.31\cdot 10^{-4}$	$2.98\cdot 10^{-1}$	$6.89\cdot 10^{-4}$	$2.13\cdot 10^{-4}$

4.3.1. A close look at one of the problems. Tables 4 and 5 provide a detailed view of the results obtained on one of the test problems—namely, the output $f = \text{Re}(S_{6\pi})$ with $n = 10d = 60$ of the low-pass filter case (see Section 4.2.1).

The results presented in these tables are the scores $R(\theta;S)$, averaged over the $M = 100$ random instances of the problem, where θ is selected using different selection criteria (along columns), and the regularity of the Matérn covariance varies or is selected automatically (along rows). The scoring rule for assessing the quality of the predictions is the SPE in Table 4 and the IS at level 95% in Table 5. (A similar table, not shown here, is presented in the SM for the CRPS.)

For comparison, Table 4 also provides the optimal values $R^\star$ obtained by direct minimization of the score (4.1). They can be used to assess the loss of predictive accuracy of the selected models, which are constructed using a limited number of observations, with respect to the best model that could have been obtained if the test data had also been used to select the parameters.

Table 4 and Table 5 support the fact that, for this particular problem, the NLL and NLL/SPE criteria are the best choices for selecting θ in terms of the SPE and the IS scores, both for a prescribed regularity ν and when ν is selected automatically (the NLL/SPE being only available for the latter case). Except for the KA criterion, however, the other selection criteria are never very far behind. Elements provided as SM show similar findings using the CRPS validation score.

Strikingly enough, for both scoring rules, the variations of the average score are much larger when the model changes than when the selection criterion changes. If a Matérn covariance function with fixed regularity is used, as is often done in practice, then the best results are obtained for all criteria (except KA) when ν takes the values $7/2$, $9/2$ and $13/2$. The values of R^* (Table 4) confirm that these

Table 5: Same as Table 4 but for averages of $R(\theta; S^{\mathrm{IS}}_{0.95})$. Using $J_n^{\mathrm{NLL/SPE}}$ gives $R(\theta; S^{\mathrm{IS}}_{0.95}) = 7.06 \cdot 10^{-2}$.

Scoring rule: $S^{\mathrm{IS}}_{0.95}$	NLL	LOO-SPE	LOO-NLPD	LOO-CRPS	KA	GCV
$\nu = 1/2$	$1.44 \cdot 10^{0}$	$1.38 \cdot 10^{0}$	$1.34 \cdot 10^{0}$	$1.48 \cdot 10^{0}$	$3.69 \cdot 10^{0}$	$1.32 \cdot 10^{0}$
$\nu = 3/2$	$2.80 \cdot 10^{-1}$	$3.05 \cdot 10^{-1}$	$2.75 \cdot 10^{-1}$	$2.88 \cdot 10^{-1}$	$3.43 \cdot 10^{0}$	$2.69 \cdot 10^{-1}$
$\nu = 5/2$	$9.30 \cdot 10^{-2}$	$9.42 \cdot 10^{-2}$	$8.55 \cdot 10^{-2}$	$9.11 \cdot 10^{-2}$	$3.36 \cdot 10^{0}$	$9.18 \cdot 10^{-2}$
$\nu = 7/2$	$6.82 \cdot 10^{-2}$	$8.82 \cdot 10^{-2}$	$8.42 \cdot 10^{-2}$	$9.10 \cdot 10^{-2}$	$3.23 \cdot 10^{0}$	$9.17 \cdot 10^{-2}$
$\nu = 9/2$	$6.50 \cdot 10^{-2}$	$9.08 \cdot 10^{-2}$	$9.30 \cdot 10^{-2}$	$9.53 \cdot 10^{-2}$	$3.14 \cdot 10^{0}$	$1.00 \cdot 10^{-1}$
$\nu = 13/2$	$6.48 \cdot 10^{-2}$	$9.95 \cdot 10^{-2}$	$9.99 \cdot 10^{-2}$	$1.03 \cdot 10^{-1}$	$3.02 \cdot 10^{0}$	$1.16 \cdot 10^{-1}$
$\nu = 25/2$	$6.77 \cdot 10^{-2}$	$1.10 \cdot 10^{-1}$	$1.08 \cdot 10^{-1}$	$1.12 \cdot 10^{-1}$	$2.90 \cdot 10^{0}$	$1.29 \cdot 10^{-1}$
$\nu = \infty$	$7.23 \cdot 10^{-2}$	$1.23 \cdot 10^{-1}$	$1.16 \cdot 10^{-1}$	$1.23 \cdot 10^{-1}$	$2.74 \cdot 10^{0}$	$1.44 \cdot 10^{-1}$
$\nu \in \{1/2, \cdots, \infty\}$	$6.70 \cdot 10^{-2}$	$1.04 \cdot 10^{-1}$	$1.06 \cdot 10^{-1}$	$1.08 \cdot 10^{-1}$	$2.78 \cdot 10^{0}$	$1.23 \cdot 10^{-1}$

are indeed the best fixed-ν models on this problem for the SPE score. Since these optimal values were not known beforehand, it is a relief to see (cf. last row of each table) that comparable performances can be achieved on this problem by selecting ν automatically.

4.3.2. Statistical analysis of the benchmark results. Tables similar to Tables 4 and 5 have been produced for all the $(4+12) \times 3 = 48$ test problems presented in Section 4.2, for the three scoring rules (SPE, CRPS and IS). We present in this section some graphical summaries and statistical analyses of these results. The individual tables for each problem are provided in the SM.

Remark 4.2. The poor performance of the KA criterion, already observed in Tables 4 and 5, is confirmed by the results (not shown) on all the other problems. We conclude that this selection criterion should not be used in practice, and exclude it from the analyses of this section in order to refine the comparison between the remaining ones.

Sensitivity analysis. We observed in Section 4.3.1 that the choice of the model—more specifically, of the regularity parameter of the Matérn covariance function—was more important than that of a particular selection criterion (excluding KA of course). To confirm this finding, a global sensitivity analysis of the logarithm of the average score, where the average is taken over the $M = 100$ instances of each problem, has been performed on each problem. The average score, for a given scoring rule, depends on two discrete factors: the selection criterion and the regularity ν of the Matérn covariance function. We present on Figure 2, for the SPE scoring rule, the Sobol' sensitivity index for the latter factor as a function of the total variance. Observe that, for the problems where the total variance is large, the Sobol index is typically very close to one, which indicates that the variability is indeed mainly explained by the choice of model. Similar conclusions hold for the other scoring rules (results not shown, see SM).

Comparison of the covariance models. Figure 3 compares the average values of $R(\theta; S^{\mathrm{SPE}})$ when ML is used on the set of GKLS problems, which have low regularities, and on the set of low-pass filter problems, which contains very smooth instances.

Observe first that the fixed-ν models rank differently on these two sets of problems, as expected considering the actual regularity of the underlying functions: low values of ν perform better on the

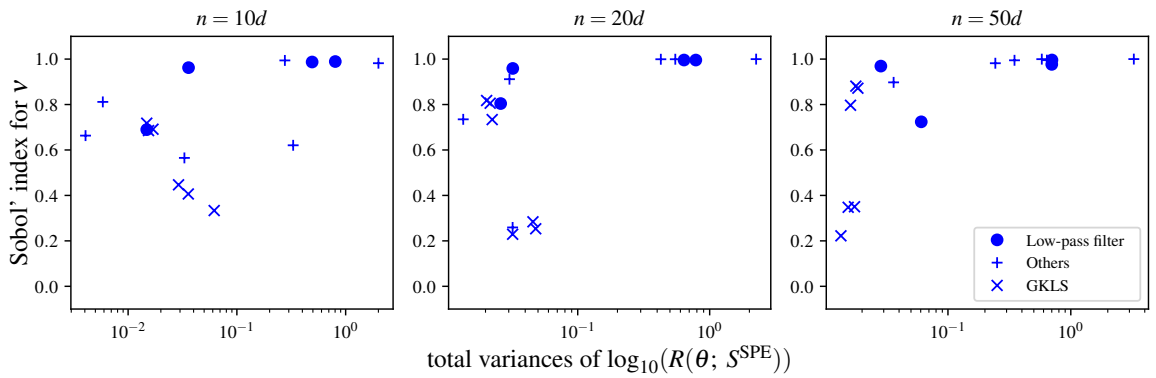

Figure 2: Parts of total variances of $\log_{10}(R)$s explained by ν for S^{SPE} using a one-factor ANOVA. Each point represents the variations of $\log_{10}(R)$ for one of the 16 problems from Section 4.2.2, split by design size, with KA and GCV excluded. The model explains almost all the variations for problems that exhibit significant fluctuations of $\log_{10}(R)$ (at the right of the figure).

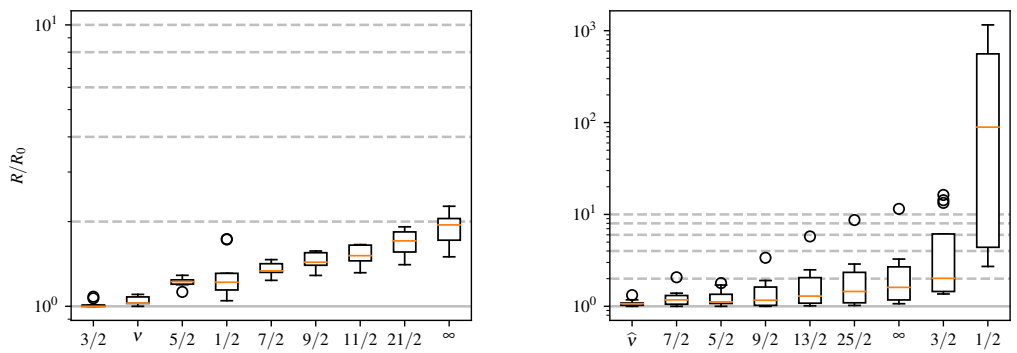

Figure 3: Box plots of R/R_0 using J_n^{NLL} as selection criterion and S^{SPE} as quality assessment criterion, for different choices of regularity. Here, R_0 stands for the best value of R on each problem (among all models). Left: All $5d$ GKLS problems. Right: All low-pass filter problems. The box plots are sorted according to their upper whisker. Grey dashed lines: $R/R_0 = 2, 4, 6, 8, 10$.

GKLS problems and worse on the low-pass filter case. Furthermore, it appears that underestimating the regularity (on the low-pass filter case) has much more severe consequences than overestimating it (on the GKLS problems) according to the SPE score, as suggested by the theoretical results of [42], [34]—see [41, Section 6] for a discussion—and [44].

Another important conclusion from Figure 3 is that very good results can be obtained by selecting the regularity parameter ν automatically, jointly with the other parameters (using the NLL criterion in this case). On the GKLS problems, the results with selected ν are not far from those of the best fixed-ν model under consideration ($\nu = 3/2$); in the low-pass filter case, they are even better than

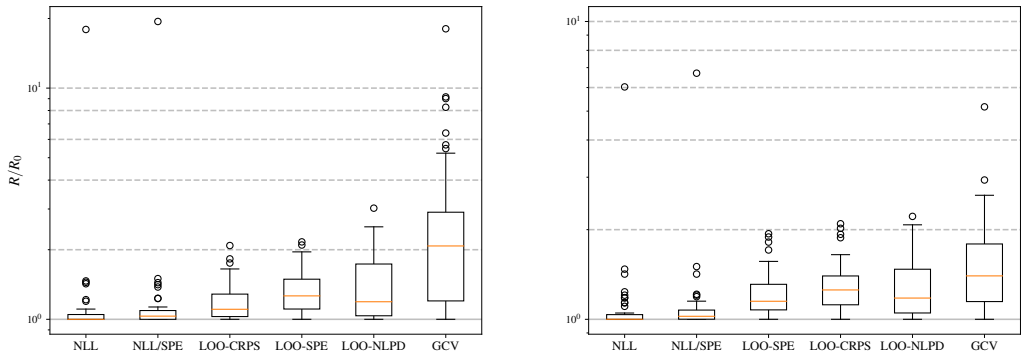

Figure 4: Box plots of R/R_0 for different selection criteria. Each box plot is constructed using all problems, with automatically selected ν, and R_0 stands for the best value of R on each problem (among all selection criteria). The left (resp. right) panel uses the S^{SPE} (resp. $S^{\mathrm{IS}}_{0.95}$) as quality assessment criterion. Sorting of box plots and horizontal lines: as in Figure 3.

those obtained with the best fixed-ν models ($\nu = 5/2$ or $7/2$). In other words, the regularity needs not be known in advance to achieve good performances, which is a very welcome practical result. This conclusion is also supported, for NLL, by the additional results provided in the SM for the other problems and for the three scoring rules.

Concerning the other selection criteria the situation is more contrasted (see SM): the automatic selection of ν using these criteria still performs very well for smooth problems, but not always, in particular with GCV, for the less regular problems of the GKLS class. This is especially true when the sample size is small ($n = 10d$).

Comparison of the selection criteria. Figure 4 compares the distributions of the average values of the SPE and IS scores for all selection criteria (except KA) on all test instances, in the case of a Matérn covariance function with automatically selected regularity. As a preliminary observation, note that for most cases the ratio R/R_0 remains under two (first horizontal dashed line), which confirms that the differences between selection criteria are much milder than those between covariance models (recall Figure 3).

A closer look at Figure 4 reveals that the rankings of criteria obtained for both scoring rules are almost identical. The ranking for the CRPS scoring rule (not shown) is the same as the one for SPE. GCV provides the worst performance for all scoring rules, followed by LOO-NLPD, while NLL dominates the ranking (for all scoring rules as well).

Remark 4.3. Observe on Figure 4 that LOO-SPE is surprisingly significantly less accurate than NLL according to S^{SPE}. More generally, choosing a scoring rule S for the LOO criterion does not guarantee the highest precision according to this particular score.

Robustness. LOO-SPE is commonly claimed in the literature (see, notably, [2]) to provide a certain degree of robustness with respect to model misspecification. According to this claim, LOO-SPE would be expected to somehow mitigate the loss of predictive accuracy with respect to likelihood-

based approaches incurred by an ill-advised choice of covariance function. Our detailed results (see SM) suggest that this effect indeed exists when the regularity is severely under-estimated (e.g., $\nu = 1/2$ for the low-pass filter problems), but is actually quite small, and should not be used to motivate the practice of setting ν to an arbitrary value. A similar effect exists for LOO-CRPS, LOO-NLPD and GCV as well. Quite surprisingly, NLL turns out to be more robust than LOO-SPE (and the other criteria) in the case of over-smoothing.

5. Conclusions. A large variety of selection criteria for Gaussian process models is available from the literature, with little theoretical or empirical guidance on how to choose the right one for applications. Our benchmark study with the Matérn family of covariance functions in the noiseless (interpolation) case indicates that the NLL selection criterion—in other words, the ML method—provides performances that are, in most situations and for all the scoring rules that were considered (SPE, CRPS and IS at 95%), better than or comparable to those of the other criteria. Considering that all the criteria tested in the study (except KA) have a similar computational complexity, this provides a strong empirical support to the ML method—which is already the *de facto* standard in most statistical software packages implementing Gaussian process interpolation.

Another important lesson learned from our benchmark study is that the choice of the family of models, and in particular of the family of covariance functions, has very often a bigger impact on performance than that of the selection criterion itself. This is especially striking when the actual function is smooth, and very irregular covariance function such as the Matérn covariance with regularity $1/2$ is used to perform Gaussian process interpolation. In such a situation, NLL is actually outperformed by other criteria such as LOO-SPE or LOO-CRPS, which thus appear to be more "robust to model misspecification". However, the small gain of performance, which is achieved by using LOO-SPE or LOO-CRPS instead of NLL in this case, is generally negligible with respect to the much larger loss induced by choosing an inappropriate covariance function in the first place.

Our final recommendation, supported by the results of the benchmark, is therefore to select, if possible, the regularity of the covariance function automatically, jointly with the other parameters, using the NLL criterion. A minimal list of candidate values for the regularity parameter should typically include $1/2$, $3/2$, $5/2$, $7/2$ and $+\infty$ (the Gaussian covariance function). Should a situation arise where a default value of ν is nevertheless needed, our recommendation would be to choose a reasonably large value such as $\nu = 7/2$, since under-smoothing has been seen to have much more severe consequences than over-smoothing. More generally, our numerical results support the fact that choosing a model carefully is important, and probably so not only in the class of Matérn covariance functions.

However, it should be kept in mind that the study focuses on cases where the number of parameters is small with respect to the number of observations (in particular, we considered zero-mean GPs with an anisotropic stationary Matérn covariance function, which have $d+2$ parameters, and we took care of having $n \gg d$). When d is large, or when the number of parameters increases, it seems to us that other selection criteria should be considered, and that the introduction of regularization terms would be required.

For future work, it would be very interesting to consider the performance of using selection criteria against a fully Bayesian approach. Another direction would be to extend this study to the case of

regression, which is also used in many applications, when dealing with stochastic simulators, for instance.

Acknowledgments. The authors thank Mona Fuhrländer and Sebastian Schöps for kindly providing the data for the "Design of a low-pass filter" example.

REFERENCES

[1] D. M. ALLEN, *The relationship between variable selection and data agumentation and a method for prediction*, Technometrics, 16 (1974), pp. 125–127.
[2] F. BACHOC, *Cross validation and maximum likelihood estimation of hyper-parameters of Gaussian processes with model misspecification*, Computational Statistics and Data Analysis, (2013).
[3] F. BACHOC, *Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case*, Bernoulli, 24 (2018), pp. 1531–1575.
[4] F. BACHOC, A. LAGNOUX, AND T. NGUYEN, *Cross-validation estimation of covariance parameters under fixed-domain asymptotics*, Journal of Multivariate Analysis, 160 (2016).
[5] M. J. BAYARRI, J. O. BERGER, R. PAULO, J. SACKS, J. A. CAFEO, J. CAVENDISH, C. H. LIN, AND J. TU, *A framework for validation of computer models*, Technometrics, 49 (2007), pp. 138–154.
[6] R. BENASSI, J. BECT, AND E. VAZQUEZ, *Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion*, in Learning and Intelligent Optimization, 5th International Conference, LION 5, C.-C. C. A., ed., Rome, Italy, January 17-21 2011, Springer, pp. 176–190.
[7] J. BERGSTRA, R. BARDENET, Y. BENGIO, AND B. KÉGL, *Algorithms for Hyper-Parameter Optimization*, in Advances in Neural Information Processing Systems 24 (NIPS 2011), vol. 24, Curran Associates, Inc., 2011.
[8] J. M. BERNARDO, *Expected information as expected utility*, The Annals of Statistics, 7 (1979), pp. 686–690.
[9] B. CALDERHEAD, M. GIROLAMI, AND N. D. LAWRENCE, *Accelerating bayesian inference over nonlinear differential equations with Gaussian processes*, in Advances in Neural Information Processing Systems 21 (NIPS 2008), 2009, pp. 217–224.
[10] C. CHEVALIER, J. BECT, D. GINSBOURGER, E. VAZQUEZ, V. PICHENY, AND Y. RICHET, *Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set*, Technometrics, 56 (2014), pp. 455–465.
[11] P. CRAVEN AND G. WAHBA, *Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation.*, Numerische Mathematik, (1979).
[12] N. CRESSIE, *Statistics for Spatial Data*, Wiley, New York, 1993.
[13] N. CRISTIANINI, J. SHAWE-TAYLOR, A. ELISSEEFF, AND J. KANDOLA, *On kernel-target alignment*, in Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01, Cambridge, MA, USA, 2001, MIT Press.
[14] C. CURRIN, T. MITCHELL, M. MORRIS, AND D. YLVISAKER, *A Bayesian approach to the design and analysis of computer experiments*, tech. report, Oak Ridge National Lab., TN (USA), 1988.
[15] C. CURRIN, T. MITCHELL, M. MORRIS, AND D. YLVISAKER, *Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments*, Journal of the American Statistical Association, 86 (1991), pp. 953–963.
[16] C. DEMAY, B. IOOSS, L. LE GRATIET, AND A. MARREL, *Model selection for Gaussian process regression: an application with highlights on the model variance validation*. 2021, https://hal.archives-ouvertes.fr/hal-03207216.
[17] L. C. W. DIXON AND G. P. SZEGÖ, *The global optimisation problem: An introduction*, in Towards Global Optimisation 2, L. C. W. Dixon and G. P. Szegö, eds., North Holland, Amsterdam, 1978, pp. 1–15.
[18] O. DUBRULE, *Cross validation of kriging in a unique neighborhood*, Journal of the International Association for Mathematical Geology, 15 (1983), pp. 687–699.
[19] G. E. FASSHAUER, *Positive definite kernels: Past, present and future*, Dolomite Res. Notes Approx., 4 (2011).
[20] G. E. FASSHAUER, F. J. HICKERNELL, E. LAMIE, P. MOKHASI, AND J. TATE, *"optimal" scaling and stable*

computation of meshfree kernel methods. Presented at the 3rd Workshop on High-Dimensional Approximation (HDA09), February 16–20, 2009, University of New South Wales, Sydney, Australia, 2009, http://math.iit.edu/~fass/FasshauerSydney.pdf.

[21] A. FORRESTER, A. SOBESTER, AND A. KEANE, *Engineering design via surrogate modelling: a practical guide*, John Wiley & Sons, 2008.

[22] M. FUHRLÄNDER AND S. SCHÖPS, *A blackbox yield estimation workflow with Gaussian process regression applied to the design of electromagnetic devices*, Journal of Mathematics in Industry, 10 (2020).

[23] M. GAVIANO, D. E. KVASOV, D. LERA, AND Y. D. SERGEYEV, *Algorithm 829: Software for generation of classes of test functions with known local and global minima for global optimization.*, ACM Trans. Math. Softw., 29 (2003), pp. 469–480.

[24] T. G. GNEITING AND A. E. RAFTERY, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association, 102 (2007), pp. 359–378.

[25] G. GOLUB, M. HEATH, AND G. WAHBA, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.

[26] GPY, *GPy: A Gaussian process framework in Python, version 1.9.9*. Available from http://github.com/SheffieldML/GPy, 2019.

[27] M. S. HANDCOCK AND M. L. STEIN, *A bayesian analysis of kriging*, Technometrics, 35 (1993), pp. 403–410.

[28] N. HANSEN, S. FINCK, R. ROS, AND A. AUGER, *Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Functions Definitions*, Research Report RR-6829, INRIA, 2009.

[29] D. JONES, M. SCHONLAU, AND W. WELCH, *Efficient global optimization of expensive black-box functions*, Journal of Global Optimization, 13 (1998), pp. 455–492.

[30] M. C. KENNEDY AND A. O'HAGAN, *Bayesian calibration of computer models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63 (2001), pp. 425–464.

[31] J. D. MARTIN AND T. W. SIMPSON, *A study on the use of kriging models to approximate deterministic computer models*, in International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2003, pp. 567–576.

[32] J. D. MARTIN AND T. W. SIMPSON, *Use of kriging models to approximate deterministic computer models*, AIAA Journal, (2005).

[33] B. MATÉRN, *Spatial Variation*, Springer-Verlag New York, 1986.

[34] F. J. NARCOWICH, J. D. WARD, AND H. WENDLAND, *Sobolev error estimates and a bernstein inequality for scattered data interpolation via radial basis functions*, Constructive Approximation, 24 (2006), pp. 175–186.

[35] C. Y. PENG AND C. F. J. WU, *On the choice of nugget in kriging modeling for deterministic computer experiments*, Journal of Computational and Graphical Statistics, 23 (2014), pp. 151–168.

[36] S. J. PETIT, J. BECT, S. DA VEIGA, P. FELIOT, AND E. VAZQUEZ, *Towards new cross-validation-based estimators for Gaussian process regression: efficient adjoint computation of gradients*, in 52èmes Journées de Statistique de la SFdS (JdS 2020), 2020.

[37] S. J. PETIT, J. BECT, P. FELIOT, AND E. VAZQUEZ, *Gaussian process model selection for computer experiments*. MASCOT PhD student 2020 Meeting. Poster presentation, 2020, https://hal-centralesupelec.archives-ouvertes.fr/hal-03018559.

[38] P. RANJAN, R. HAYNES, AND R. KARSTEN, *A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data*, Technometrics, 53 (2011), pp. 366–378.

[39] C. E. RASMUSSEN AND C. K. I. WILLIAMS, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, USA, 2006.

[40] T. J. SANTNER, B. J. WILLIAMS, AND W. I. NOTZ, *The Design and Analysis of Computer Experiments.*, Springer series in statistics, Springer, 2003.

[41] M. SCHEURER, R. SCHABACK, AND M. SCHLATHER, *Interpolation of spatial data – a stochastic or a deterministic problem?*, European Journal of Applied Mathematics, 24 (2013), p. 601–629.

[42] M. STEIN, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Series in Statistics, Springer New York, 1999.

[43] S. SUNDARARAJAN AND S. S. KEERTHI, *Predictive approaches for choosing hyperparameters in Gaussian pro-*

cesses, Neural Computation, 13 (2001), pp. 1103–1118.

[44] R. TUO AND W. WANG, *Kriging prediction with isotropic matern correlations: Robustness and experimental designs*, Journal of Machine Learning Research, 21 (2020), pp. 1–38.

[45] G. WAHBA, *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[46] G. WAHBA AND J. WENDELBERGER, *Some new mathematical methods for variational objective analysis using splines and cross validation*, Monthly Weather Review, 108 (1980), pp. 1122–1143.

[47] H. WENDLAND, *Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree*, Advances in Computational Mathematics, 4 (1995), pp. 389–396.

[48] R. WILKINSON, *Accelerating ABC methods using Gaussian processes*, in Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS 2014), vol. 33 of Proceedings of Machine Learning Research, PMLR, 2014, pp. 1015–1023.

[49] B. A. WORLEY, *Deterministic uncertainty analysis*, Tech. Report ORNL–6428, Oak Ridge National Laboratory, TN, USA, 1987.

[50] Z. YING, *Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process*, J. Multivar. Anal., 36 (1991).

[51] H. ZHANG, *Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics*, Journal of the American Statistical Association, 99 (2004), pp. 250–261.

[52] H. ZHANG AND Y. WANG, *Kriging and cross-validation for massive spatial data*, Environmetrics, 21 (2010), pp. 290–304.