SIAM/ASA J. Uncertainty Quantification
Vol. 6, No. 3, pp. 991–1018

Warped Gaussian Processes and Derivative-Based Sequential Designs for Functions with Heterogeneous Variations*

Sébastien Marmin†, David Ginsbourger‡, Jean Baccou§, and Jacques Liandrat¶

Abstract. Gaussian process (GP) models have become popular for approximating and exploring nonlinear systems using scarce input/output samples and prior hypotheses done through mean and covariance functions. While it is common to make stationarity assumptions and use variance-based criteria for exploration, in realistic cases it is not rare that systems under study exhibit a heterogeneous behavior depending on regions of the parameter space. We consider a class of problems where high variations occur along unknown noncanonical directions and we tackle the problem of accommodating nonstationarity from two angles. First we define a novel class of covariances (WaMI-GP) that simultaneously generalizes kernels of multiple index and of tensorized warped GPs, and second, we introduce derivative-based sampling criteria dedicated to the exploration of high-variation regions. The novel GP class is investigated through both mathematical analysis and numerical experiments, and it is shown that it allows encoding much expressiveness while keeping the number of parameters to be inferred moderate. Criteria and models are compared on a mechanical test case from safety studies conducted by IRSN. On this application some of the proposed criteria outperform usual variance-based criteria in the case of a stationary GP model; however, variance-based criteria with WaMI-GP perform even better. Our method is also compared with the treed Gaussian processes (TGP) on this application and on a NASA test case. In the IRSN application, WaMI-GP dominates TGP in static and sequential settings. In the NASA application, while TGP clearly dominates in the static case, for small designs it is outperformed by WaMI-GP in the sequential setup.

Key words. nonstationary kernels, infill sampling criteria, computer experiments

AMS subject classifications. 97K80, 62L05, 60G15, 46N30

DOI. 10.1137/17M1129179

1. Introduction.

Many systems abruptly change regime: in materials sciences and fluid mechanics, with percolation in porous media, in epidemiology, with outbreak of a pathogen depending on population characteristics, in thermodynamics with phase transition, etc. This situation is also encountered in nuclear safety analysis where, for instance, slight variations in input parameters of computer codes may strongly impact responses quantifying system

*Received by the editors May 8, 2017; accepted for publication (in revised form) May 7, 2018; published electronically July 3, 2018.

http://www.siam.org/journals/juq/6-3/M112917.html

†Institut de Radioprotection et de Sûreté Nucléaire, Cadarache, France, and Institute of Mathematical Statistics and Actuarial Science, Department of Mathematics and Statistics, University of Bern, Bern, Switzerland (sebastien.marmin@irsn.fr).

‡Uncertainty Quantification and Optimal Design Group, Idiap Research Institute, Martigny, Switzerland, and Institute of Mathematical Statistics and Actuarial Science, Department of Mathematics and Statistics, University of Bern, Bern, Switzerland (ginsbourger@idiap.ch).

§Laboratoire de Micromécanique et d'Intégrité des Structures, IRSN-CNRS-UM, Cadarache, France (jean.baccou@irsn.fr).

¶Aix Marseille Université, CNRS, Centrale Marseille, I2M, UMR7353, 13451 Marseille, France (jacques.liandrat@centrale-marseille.fr).

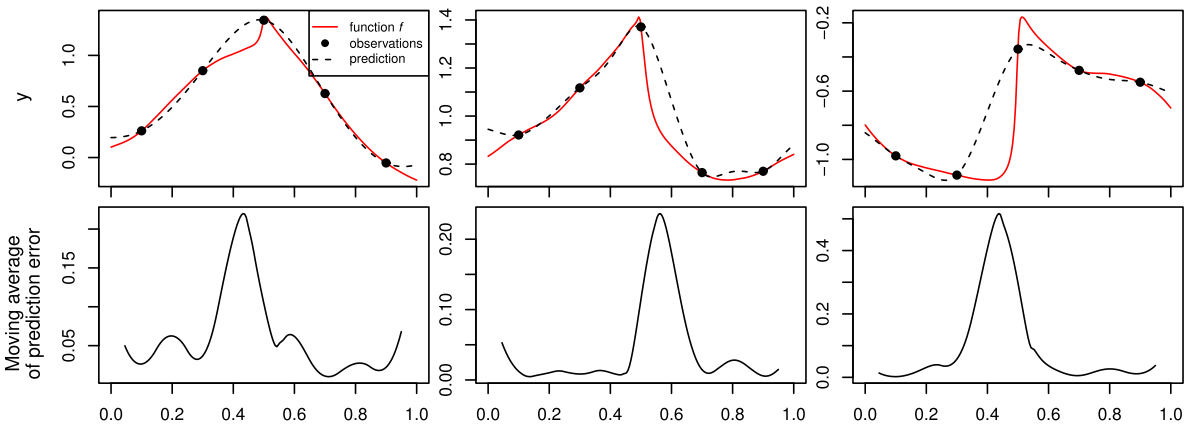

Figure 1. *Illustration of the concentration of prediction errors around high-variation regions. Top: functions with heterogeneous variations (sample paths of a nonstationary GP) interpolated based on five evaluations. Predictions from a GP model with mean zero and stationary covariance of type Matérn* $\nu = \frac{5}{2}$*. Bottom: absolute differences between predictions and true values, averaged on a moving window of width* $\frac{1}{10}$*.*

safety due to a steep transition between competing mechanical phenomena. Let us focus on one real-valued response of some system with respect to d variables, formally a function $f : \boldsymbol{x} \in D \subset \mathbb{R}^d \to f(\boldsymbol{x}) \in \mathbb{R}$. For differentiable f's, abrupt changes of regime are reflected in the evolving magnitude of the gradient norm depending on regions of D or, to take one alternative viewpoint, by spatially varying main local frequencies. Here we informally refer to f's exhibiting such behavior as "functions with heterogeneous variations." As illustrated in Figure 1, regions with abrupt variations of the response can lead to increased prediction errors. Allocating more evaluations in these regions is a natural idea. However, practitioners often only guess the existence of such heterogeneities without much precise information regarding their location, shape, or orientation. If the function f is expensive to evaluate it is reasonable to appeal to modeling and sampling approaches that allow for uncovering such regions based on data and, ultimately, for better approximating f. In *sequential* design of experiments, the choice of evaluation points is typically guided by a (cheap) surrogate model of f. Often surrogate model predictions come with prediction uncertainties, and sampling criteria rely upon them in order to determine the next evaluation point(s). Evaluations of f at points deemed most promising and surrogate model updates are then repeated until some stopping condition is met, e.g., depletion of the evaluation budget.

Gaussian process (GP) models have become a popular surrogate class, especially for the design and analysis of computer experiments (see, e.g., [37], [21], and [39]). GP modeling consists in assuming that the unknown objective function f is a priori a sample path of a GP, denoted by Y, indexed by the source space of f, leading to posterior distributions when taking function evaluation results into account. The quality of the model depends both on the evaluation points and on the adequacy between the prior GP (parametrized by its mean and covariance functions) and the function to be predicted. Adapting the prior covariance of Y to specific classes of objective functions f has inspired several contributions in GP modeling. For example, for objective functions with a better representation in polar coordinate, [26] proposes

GP models that incorporate the geometry of the disk. Similarly, appropriate prior covariances exist for functions known to satisfy degeneracies such as symmetries or harmonicity [13], and for functions with a sparse ANOVA decomposition [11, 14]. In the absence of such specific prior assumption on f, it is common to take stationary covariance functions [43]. Consider a constant-mean GP model; then a stationary prior covariance means that the distribution of outputs $(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'})^\top$ solely depends on $(\boldsymbol{x}, \boldsymbol{x}') \in D^2$ via the difference $\boldsymbol{x} - \boldsymbol{x}'$. Among stationary kernels, the Matérn class is quite popular as it allows one to tune the order of (almost sure) differentiability of associated GP realizations. Note that both tensor product Matérn kernels (e.g., in [36]) and their radial counterparts (such as in [33]) have been used. For sequential settings it is interesting to keep in mind, however, that a number of properties including stationarity vanish when conditioning on data, as discussed notably in [35].

Yet, when f is known to possess heterogeneous variations, it is sensible to consider nonstationary prior covariances that account for this property. Various nonstationary GP models were proposed in the literature, notably convolution methods (see [25, 12]) and input space warping approaches [38]. In warping approaches, nonstationarity comes from the chaining of a GP with a warping function. A strongly consistent approach for estimating deformations of a bivariate isotropic GP from dense evaluations of a single (deformed) realization is provided by [1]. In contrast, challenges considered here rather call for estimating nonnecessarily bijective warpings from scarce evaluations in order to build nonstationary surrogates for functions with arbitrary d-dimensional source space. Gibbs [12] tackled this problem using parametric warpings relying on linear combinations of basis functions. Following this idea, Xiong et al. [47] drastically reduced the number of parameters by taking tensor products of univariate warpings.

An additional popular model, the treed Gaussian process (TGP) [15], consists in partitioning the input space D into parallelepipeds on which individual GPs are defined and then combined. While this method is very flexible and allows, by construction, accounting for heterogeneous variations, it requires appealing to the machinery of posterior sampling and does not enjoy the convenient analytical tractability of the plain GP approach. In the context of small data sets and heterogeneous variations driven by unknown directions, there is a need of GP surrogate models that enjoy the sparsity of the axial warping of Xiong et al. while keeping nice flexibility properties of TGP or the Gibbs approach without relying on canonical axes. The last point notably refers to *single index models* (SIM) such as GP-SIM [19] and more generally to *multiple index models* (MIM) [46]. We propose a kernel class inspired by these considerations, as detailed in the next sections.

From a different perspective, GPs have been used for sequential design of computer experiments, notably with variance-based sampling criteria like the mean squared error (MSE) and integrated MSE (IMSE) that allow allocating evaluations to unexplored regions. While strategies based on such criteria tend to fill the design space [44], in cases where covariance parameters are not reestimated it is done nonadaptively, as the MSE does not depend on observations but solely on the location of points. In contrast, GP-based adaptive criteria have been tackled for estimating target regions such as contour lines, excursion sets, and related [45, 32, 29, 2, 4]. On a different note, adaptive design criteria have been used for global optimization (see notably [23, 21, 42]). In particular, input warping has been recently shown to improve Bayesian optimization in nonstationary cases [41].

Our contributions concern the problem of learning functions with heterogeneous variations along unknown directions both from the modeling and the sequential design point of view. First we introduce the warped multiple index (WaMI) GP model, relying on a new family of nonstationary covariance kernels that combines features from multiple index GPs and tensorized warpings. A nice aspect of this kernel family is that the number of hyperparameters increases affinely with d (with slope 1). Besides this, the model can incorporate any orientation of heterogeneous variations. Regarding the sequential design aspect, we develop targeted criteria based on GP gradient norms that make a trade-off between space-fillingness and intensifying exploration in high-variation regions.

We apply these contributions on functions from two mechanical engineering case studies. The first test case stems from numerical simulations of fracture dynamics arising in risk studies at the French Institute for Radioprotection and Nuclear Safety (IRSN). On this test case, the WaMI-GP model outperforms both a stationary GP model and a TGP model in static prediction from a class of initial designs. Moreover the same test case is used to compare performances of sequential design strategies by varying both criteria (MSE and IMSE versus introduced derivative-based criteria) and surrogate models. Best results are obtained using WaMI-GP (and to a lesser extent, TGP) combined with classical variance-based criteria, followed by a stationary GP model combined with one of the proposed gradient-based criteria. In the second test case, we study a three-dimensional fluid dynamics application from NASA that was used in seminal article about TGP. In this test case, TGP remains the best model at fixed space-filling design of experiments, but when both surrogates are combined with MSE-based sequential design, WaMI-GP catches up TGP by successfully detecting high-variation regions and leading to better prediction performance.

The paper is organized as follows. Section 2 is devoted to an overview on GP models with foci on nonstationarity and on GP-driven sequential design of experiments. Then we introduce and investigate the WaMI-GP model in section 3, followed in section 4 by several proposals of derivative-based criteria for exploring high-variation regions. Finally, experimental comparisons with classical approaches based on the two engineering test cases are presented in section 5.

2. State of the art.

2.1. Stochastic modeling for the emulation of computer experiments.

2.1.1. Overview of Gaussian process modeling basics. In GP modeling, one assumes that the objective function is a realization of a Gaussian random field $Y \sim \mathcal{GP}(m(\cdot), c(\cdot,\cdot))$ indexed by D, specified in distribution by its mean and covariance functions $m(\cdot)$, $c(\cdot,\cdot)$. These functions are typically taken among some parametric families and parameters are either estimated and plugged in or treated as random variables in the full Bayesian framework. Let us denote $m = m_{\boldsymbol{\theta}_1}$ and $c = c_{\boldsymbol{\theta}_2}$, with an overall *hyperparameter* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. In this article, we adopt an empirical Bayes viewpoint, i.e., that the Bayesian paradigm is used when seeing f as a random element with a GP prior, but the hyperparameter $\boldsymbol{\theta}$ is treated as deterministic even if it is estimated from data. In other words, given $\boldsymbol{\theta}$, $\forall\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in D$, the corresponding response vector $(Y_{\boldsymbol{x}_1}, \ldots, Y_{\boldsymbol{x}_n})^\top$ (standing for values of the objective functions at those points) is a priori distributed as a multivariate normal distribution $\mathcal{N}(\mathfrak{m}_n, C)$, with

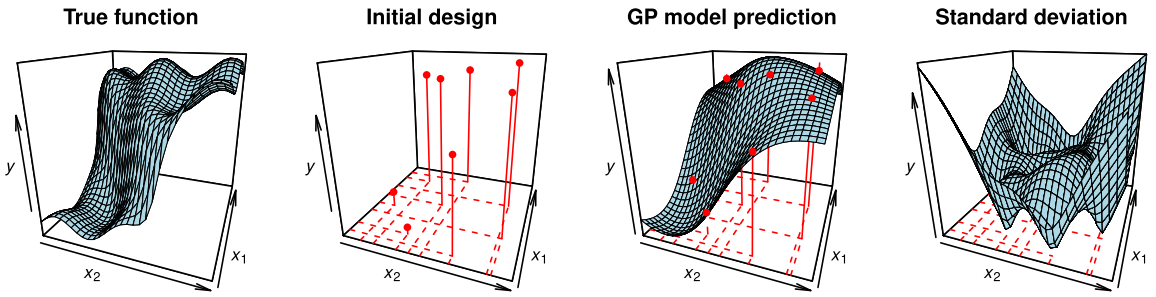

Figure 2. *Stationary GP modeling of a toy function* (3)*. From left to right: f; locations and responses of eight initial evaluations; posterior GP mean; posterior GP standard deviation.*

$\mathfrak{m}_n = (m_{\boldsymbol{\theta}_1}(\boldsymbol{x}_i))_{1\le i\le n}^\top$ and $C = (c_{\boldsymbol{\theta}_2}(\boldsymbol{x}_i, \boldsymbol{x}_j))_{1\le i,j\le n}$. Estimators for $\boldsymbol{\theta}$ can notably be defined by cross-validation minimization or by maximum likelihood (see [27]) and then simply plugged in. Details about mean and covariance parameter estimation can be found in [36]. Note that even though we stick to the empirical Bayes setup for simplicity, most results presented throughout the paper could be extended naturally to the full Bayesian framework.

Now given n arbitrary points $\boldsymbol{x}_1, \dots \boldsymbol{x}_n \in D$ and observed values $(y_1, \dots, y_n)$ of Y at those points, the so-called kriging formulas and the underlying posterior GP model are obtained by conditioning Y on the event $\mathcal{A}_n = \{Y_{\boldsymbol{x}_1} = y_1, \dots, Y_{\boldsymbol{x}_n} = y_n\}$:

$$m_n(\boldsymbol{x}) = \mathbb{E}\left(Y_{\boldsymbol{x}} \middle| \mathcal{A}_n\right) = m_{\boldsymbol{\theta}_1}(\boldsymbol{x}) + \mathfrak{c}_n^\top C^{-1}(\boldsymbol{y}_{1:n} - \mathfrak{m}_n) \text{ and} \tag{1}$$

$$c_n(\boldsymbol{x}, \boldsymbol{x}') = \operatorname{cov}\left(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'} \middle| \mathcal{A}_n\right) = c_{\boldsymbol{\theta}_2}(\boldsymbol{x}, \boldsymbol{x}') - \mathfrak{c}_n^\top C^{-1}\mathfrak{c}_n, \tag{2}$$

where $\boldsymbol{y}_{1:n} = (y_i)_{1\le i\le n}^\top$, $\mathfrak{c}_n = (c_{\boldsymbol{\theta}_2}(\boldsymbol{x}, \boldsymbol{x}_i))_{1\le i\le n}^\top$ and C is assumed nonsingular here. Figure 2 shows a bivariate stationary GP model with

$$f : \boldsymbol{x} \in [0,1]^2 \to \left(\sin(15x_1) + \cos(10x_2)\right)/5 + \arctan\left(10(x_1 + x_2) - 15/2\right). \tag{3}$$

A model is built from eight observations at a Latin hypercube design (optimized with a maximin distance criterion; see, e.g., [10]). While trends are parametrized by basis functions coefficients in the case of *universal kriging*, here we focus mostly on the role of the covariance and the trend is typically taken as a constant (estimated in the *ordinary kriging* setting, as discussed in [36]). The covariance kernel used in Figure 2 is an anisotropic stationary Matérn with $\nu = 5/2$ and $\boldsymbol{\theta} \in (\mathbb{R}_+\backslash\{0\})^{d+1}$, i.e., $c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = \theta_{d+1}(1 + \sqrt{5}h + \frac{5}{3}h^2)\exp(-\sqrt{5}h)$, where $h = \sqrt{\sum_{i=1}^d \frac{(x_i - x_i')^2}{\theta_i^2}}$ (see [43, 33]). In this example the function f has heterogeneous variations across the input space in the sense that there is a steep region in a band around the line of equation $x_2 = 3/4 - x_1$. Localized features question the choice of a stationary covariance, and actually a nonstationary model may improve the model by fitting the heterogeneous behavior of f.

2.1.2. Nonstationary approaches: From warped GP to TGP. There are several approaches to inject prior knowledge about spatial-dependency. A trivial way is vertical scaling: $c(\boldsymbol{x}, \boldsymbol{x}') = \sigma(\boldsymbol{x})\sigma(\boldsymbol{x}')R(\boldsymbol{x}, \boldsymbol{x}')$, where R is a correlation function and σ a nonnegative (nonconstant) function is a valid nonstationary covariance (see, e.g., [22]). Moreover, Paciorek and

Schervish [25] address the issue by extending a convolution method proposed in [12] from a squared exponential kernel to any covariance structure. Warping stationary GPs for creating nonstationary GPs is also a common method (see, e.g., [38]). In this approach, called the nonlinear map method, the nonstationary covariance c is derived from $\boldsymbol{x} \to Y_{\boldsymbol{x}} = Z_{T(\boldsymbol{x})}$, with Z a (stationary) GP of covariance $k(\cdot,\cdot)$ on $\mathbb{R}^p \times \mathbb{R}^p$ and T a function from D to $\mathbb{R}^p$. The covariance of Y is then $\forall \boldsymbol{x}, \boldsymbol{x}' \in D$, $c(\boldsymbol{x}, \boldsymbol{x}') = k(T(\boldsymbol{x}), T(\boldsymbol{x}'))$.

The nonlinear map method's flexibility turns out to be challenging in practice as it requires one to estimate from data a warping T among the set of all injections on D. A first restriction, implicitly assumed in almost all applications, is to consider only continuous bijections. The estimation of T is often simplified to a finite-dimensional problem, taking $T = T_{\boldsymbol{\tau}}$, with $\boldsymbol{\tau}$ a parameter vector. Gibbs' method [12], for example, formulates $T_{\boldsymbol{\tau}}$ as a multidimensional integral of nonnegative density functions that ensures continuity (and bijectivity in the case of positive densities; see also Appendix A for more details and an illustration). In this method, keeping the same level of spatial precision, say, r basis functions for each direction, the number of weights is dr^d. This reduces the applicability of the method in contexts with drastically limited numbers of evaluations. For this reason, further work focused on reducing the number of parameters while preserving some flexibility.

In the axial warping method of [47], it is assumed that for $\boldsymbol{x} \in D$, $T(\boldsymbol{x}) = (T_i(x_i))_{1 \le i \le d}^\top$, with $(T_i)_{1 \le i \le d}$ continuous univariate bijections. The functions T_i, $i = 1, \ldots, d$, are taken piecewise second degree polynomials, with differentiability constraints and equally spaced nodes. In Figure 3 we display the results of applying this method to the synthetic example (3), along with the estimated axial warping densities. In some situations, warping only along canonical axes can be questioned. For instance, if the expected, or "real," warping is of the form $T(\boldsymbol{x}) = \boldsymbol{x} + T_1(\boldsymbol{x}^\top \boldsymbol{u})\boldsymbol{u}$, with $\boldsymbol{u}$ an arbitrary noncanonical direction in $\mathbb{R}^d$, an axial warping cannot incorporate that orientation. Although this warping is simple, and potentially useful in many applications, the general Gibbs approach needs a lot of parameters to approximate T. In Xiong et al. the number of parameters is reduced but this simplification appears to be too rigid in some applications.

Using a TGP is another strategy for modeling functions with heterogeneous variations. This method is based on partitioning the input space. Independent GP models defined over

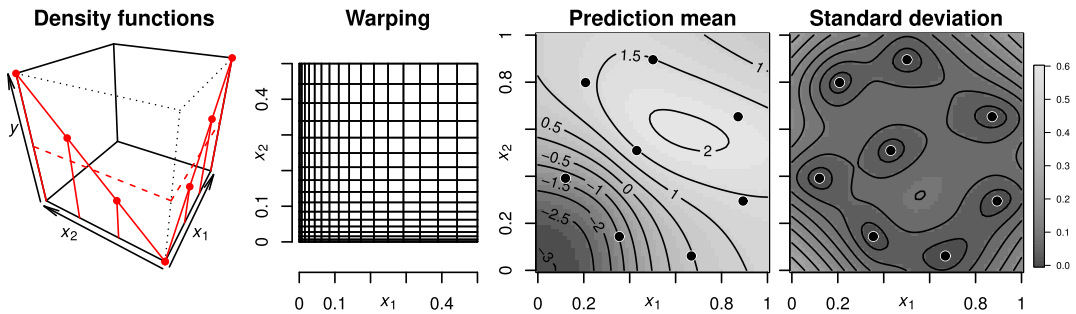

Figure 3. *GP model with axial warping, applied to the example fuction Equation (3). From left to right: estimated warping density functions for the axes; corresponding surface warping of* $[0,1]^2$ *(represented by deformation of a* 10×10 *regular orthogonal grid); prediction mean and standard deviation.*

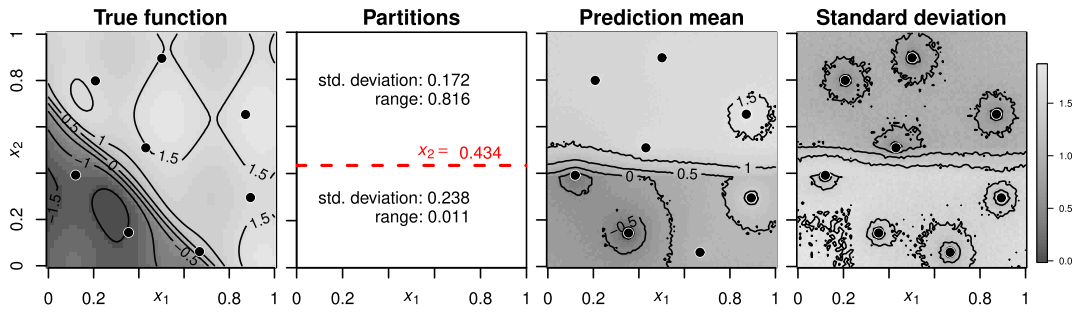

Figure 4. *Bayesian treed GP model. From left to right: the objective function with an initial design; a sketch of the input space partition, where the red line divides the space into two regions with different range and scale parameters for different GP models; the TGP prediction mean and standard deviation.*

the different subregions are combined, allowing highly heterogeneous behavior across the input space. A strength of this method is that partitions are automatically determined based on data. The discontinuities resulting from partitioning could sound like a drawback, but Bayesian averaging mitigates their detrimental effects in terms of prediction error. Figure 4 shows the application on the toy function obtained with the R package *tgp* [16, 20].

We observe in this example that the TGP model leads to a partition of the input space in two zones. Indeed the region $x_2 > 0.434$ appears to have fewer variations than the region $x_2 \leq 0.434$. The partitions are implemented to be defined in terms of the canonical axes. Here, it can be expected that partitioning the space with respect to the line $x_1 + x_2 = 3/4$ could constitute an improvement, because it takes into account the transition region. If the direction of the line were known a priori, TGP could accomplish an estimation of its position by putting $x_3 = x_1 + x_2$ in as a third predictor and allowing partitioning only along that dimension.

2.1.3. Dimension reduction with the multiple index model. In most anisotropic GP models, the dimension of $\boldsymbol{\theta}$ increases rapidly with d. Geometric anisotropies in dimension d require one to parametrize a rotation ($d(d-1)/2$ parameters) and length-scale parameters for each relevant direction (see, e.g., [33, p. 10] for a presentation of the squared exponential anisotropic GP). To avoid this quadratic increase in the number of parameters, one can consider alternatives like the SIM [3]. In GP modeling, Gaussian process-SIM (GP-SIM [6, 19]) has for prior covariance $c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\boldsymbol{\beta}}(\boldsymbol{a}^\top \boldsymbol{x}, \boldsymbol{a}^\top \boldsymbol{x}')$, where $k_{\boldsymbol{\beta}}$ is a covariance kernel over $\mathbb{R} \times \mathbb{R}$ and is hence parametrized by a vector $\boldsymbol{\beta}$ and $\boldsymbol{a} \in \mathbb{R}^d$. In an empirical Bayesian setting, this model produces constant predictions in all hyperplanes orthogonal to $\boldsymbol{a}$. Relaxing this constraint, the MIM is an extension proposed by [46]. It extends the scalar product to a matrix product, resulting in the GP framework to covariance kernels of the form

$$c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\boldsymbol{\beta}}\left(A\boldsymbol{x}, A\boldsymbol{x}'\right) \tag{4}$$

with $q \in \mathbb{N}\backslash\{0\}$, A a $q \times d$ matrix, and $k_{\boldsymbol{\beta}}$ a covariance kernel on $\mathbb{R}^q$ parametrized by $\boldsymbol{\beta}$. With this covariance function, the dimension of $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{a}\}$ increases affinely in d with slope q. An important advantage of MIM is its neutrality toward the canonical axes: any invertible linear pretreatment of the data intrinsically leads to the same estimation problem, which is not the case for many models relying on the canonical axes.

2.2. Sampling: Principle and classic criteria. Once a GP model has been built from an initial, e.g., space-filling, design of N_{ini} evaluations (commonly used designs are optimized Latin hypercube sample (LHS) designs and minimax-distance designs; see [30] for an overview), the sequential design itself involves a loop over n, the current number of evaluations $n = N_{\text{ini}} + 1, \ldots, N$. Sequential sampling is typically driven by successive optimizations of infill criteria J_n, evaluations of the objective function at resulting points, and model updates. More precisely, in fully sequential settings, the next evaluation $\boldsymbol{x}_{n+1}$ is selected as a point maximizing a criterion J_n:

$$\boldsymbol{x}_{n+1} \in \operatorname*{argmax}_{\boldsymbol{x} \in D} J_n(\boldsymbol{x}). \tag{5}$$

A criterion depends on past evaluations and is defined in terms of the mean $m_n(\cdot)$ and the covariance $c_n(\cdot, \cdot)$ of the GP at step n.

A common principle to several design classes is to sequentially evaluate f at points chosen with the aim to maximally reduce GP prediction uncertainty, often defined with the help of posterior variance functions. MSE and IMSE criteria are based on this rationale. These criteria focus on zones where prediction variance is the highest (or where its integral is susceptible of being maximally reduced in expectation), i.e., where the function is still (relatively) poorly explored in the sense of the GP model. For IMSE, the aim is to reduce the integral of the MSE over the whole domain D. Thus minimizing the IMSE corresponds to looking for a point $\boldsymbol{x}$ minimizing the integral of the future MSE if $\boldsymbol{x}$ is added:

$$J_n^{\text{IMSE}}(\boldsymbol{x}) = \int_{\boldsymbol{u} \in D} c_{n,\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{u})\, \mathrm{d}\boldsymbol{u}, \tag{6}$$

with $c_{n,\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{u}) = \operatorname{var}(Y_{\boldsymbol{u}} | \mathcal{A}_{n,\boldsymbol{x}})$, $\mathcal{A}_{n,\boldsymbol{x}} = \mathcal{A}_n \cup \{(\boldsymbol{x}, m_n(\boldsymbol{x}))\}$. The term $c_{n,\boldsymbol{x}}$ can theoretically be obtained using the kriging formula (2), but substantial computational savings are made using an "update formula"; see [5] for details.

Figure 5 illustrates the first step of a sequential sampling procedure on our running toy function under the three previously recalled models (stationary, nonstationary with axial

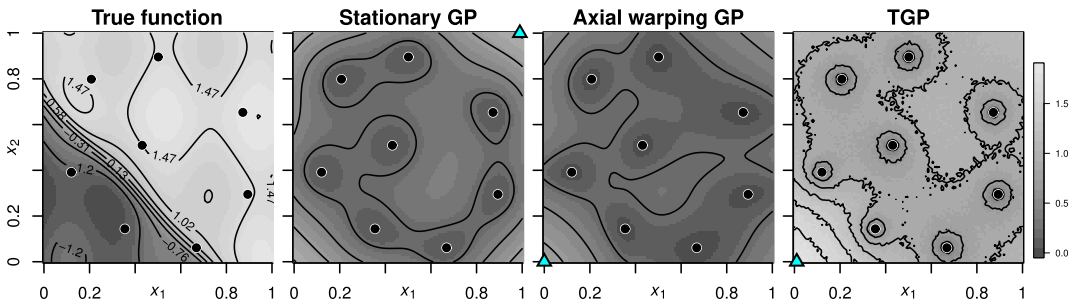

Figure 5. *MSE criterion in the first step of a sequential sampling procedure of* $f : \boldsymbol{x} \in [0,1]^2 \to \frac{\sin(15x_1)+\cos(10x_2)}{5} + \arctan(\frac{20(x_1+x_2)-15}{2})$ *(left), according to the model, stationary anisotropic GP, Xiong's axial warping GP, and treed GP. The blue triangles represent the maximum of MSE, i.e., the next evaluation point.*

warpings, and TGP). Three MSE maximizers that depend on the model can be seen as proposals for the next evaluations.

Nonstationary modeling appears to be a promising choice to tackle the problem of designing experiments in the case of objective functions exhibiting heterogeneous variations. However, as mentioned in subsection 2.1.2, existing methods appear to not be fully satisfactory for reducing evaluation budget in some situations.

3. WaMI-GP: A nonstationary multiple index model. We now present WaMI-GP, a novel GP model dedicated to multivariate functions with heterogeneous variations along noncanonical axes. This class of models involves a number of parameters describing the axes as well as the deformations coming into play; however, their cardinality is kept moderate thanks to the tensorial nature of the involved warping, i.e., by writing the warping as a vector of univariate deformations. In this section we introduce the model, illustrate its flexibility, and prove some of its important properties. Finally we show how WaMI outperforms other considered GP classes on our running example both statically and when combined with a state-of-the-art sequential design of experiments approach.

3.1. Formulation of the WaMI covariance. We focus on the covariance kernel of the proposed GP class as, without loss of generality, the GP mean is assumed constant.

Definition 3.1 (WaMI kernel: combining deformations and multiple index modeling). *Let $q \in \mathbb{N}\backslash\{0\}$, $A \in \mathbb{R}^{q\times d}$, $T_i(\cdot,\boldsymbol{\tau}_i) : \mathbb{R} \mapsto \mathbb{R}$ be functions parametrized by $\boldsymbol{\tau}_i$ $(i = 1,\dots,q)$ and $k_{\boldsymbol{\beta}}$ be a positive definite kernel on $\mathbb{R}^q$ parametrized by $\boldsymbol{\beta}$. Assuming that the parametric form of the T_i's is given and denoting by $\boldsymbol{\theta}$ a hyperparameter consisting of A, the $\boldsymbol{\tau}_i$'s, and $\boldsymbol{\beta}$, we define the associated WaMI kernel on D by*

$$c_{\boldsymbol{\theta}} : (\boldsymbol{x},\boldsymbol{x}') \in D\times D \to c_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{x}') = k_{\boldsymbol{\beta}}\left(T(A\boldsymbol{x}),T(A\boldsymbol{x}')\right) \tag{7}$$

with $T(\boldsymbol{u}) = (T_i(u_i;\boldsymbol{\tau}_i))_{1\le i\le q}$.

The symmetry and positive definiteness (in the wide sense) of the WaMI kernel are inherited from the basis kernel $k_{\boldsymbol{\beta}}$ through the overall warping, consisting of A and the *univariate deformations* (or *univariate warpings*) T_i. More is established below in subsection 3.3 after some examples. Before we get there, let us get a feeling about how the WaMI kernel fits into the axial warping and MIM frameworks:

1. As an axial warping method, WaMI allows noncanonical directions for orientation of the univariate deformations by acting on the input space via a linear map with matrix A. Note that this kernel also accommodates dimension reduction (and thus reducing the number of axial warpings) in case $q < d$.
2. From the MIM perspective, we introduce nonstationarity into the covariance by applying nonlinear deformations to the result of each scalar product $\boldsymbol{a}_i^\top\boldsymbol{x}$ (with $A = [\boldsymbol{a}_1,\dots,\boldsymbol{a}_q]^\top$).

Naturally, it is possible to take identity T_i's for one to several dimensions, hence reducing the number of deformations and also of covariance parameters. Besides this, the class can theoretically be generalized to cases where the warpings are not scalar but rather defined on subspaces of $\mathbb{R}^q$. With this parametrization the total number of parameters is $qd + \#\boldsymbol{\beta} + \sum_{i=1}^q \#\boldsymbol{\tau}_i$, where $\#\boldsymbol{\alpha}$ stands for the dimensionality of $\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is an arbitrary parameter.

Let us now focus on the identifiability of the covariance parameters. We first consider the case of linear T_i's and stationary GP prior for $k_{\boldsymbol{\beta}}$, written as $k_{\boldsymbol{\beta}}(\boldsymbol{u}, \boldsymbol{u}') = \beta_0 R(B^{-1}(\boldsymbol{u} - \boldsymbol{u}'))$ (where R is a positive definite kernel on $\mathbb{R}^q$, $\beta_0 > 0$ is the variance, and B is an invertible $d \times d$ matrix), and $q = d$. We see from (7) why the model is overparameterized as we identify $B^{-1}A$ and not B and A separately. To address this issue, one can restrict A to be a rotation matrix (parametrized by angles) and B^{-1} to be a diagonal matrix. This reduces the count of parameters to

$$d(d-1)/2 + \underbrace{d}_{\#\mathrm{diag}B} + \underbrace{1}_{\#\beta_0} + \sum_{i=1}^{q} \#\boldsymbol{\tau}_i. \tag{8}$$

Another identifiability issue appears in case of an isotropic R: $B^{-1}A$ is estimated up to an orthogonal matrix. This means the parameter space for A and B can be further reduced, with restrictions on the sign of the diagonal terms of B and on the interval of the angles of A. However, these additional restrictions do not reduce the number of parameters and the values of A and B are not of direct interest. What matters here is the overall warping $\boldsymbol{x} \to BT(A\boldsymbol{x})$ up to a composition with an isometry.

When the T_i's are not assumed linear, there is in general no redundancy between A and the parameters of $k_{\boldsymbol{\beta}}$. However, to limit the number of parameters, in the rest of the article we keep A and B as rotation and diagonal matrices, respectively.

For example, fixing the T_i's to some prescribed linear or nonlinear warpings, a two-dimensional instance of the WaMI kernel class of Definition 3.1 can be obtained by putting

$$A = \begin{pmatrix} \cos(\theta_0) & -\sin(\theta_0) \\ \sin(\theta_0) & \cos(\theta_0) \end{pmatrix} \quad \text{and} \quad k_{\boldsymbol{\beta}}(\boldsymbol{u}, \boldsymbol{u}') = \beta_0 k_{\text{Matérn}}\left(\frac{1}{\beta_1}\sqrt{\boldsymbol{u}^\top \boldsymbol{u}'}\right) \tag{9}$$

with $\boldsymbol{\theta} = (\theta_0, \beta_0, \beta_1, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$, $\beta_0, \beta_1 > 0$, and $k_{\text{Matérn}}$ is the Matérn kernel with smoothness parameter $\nu = 5/2$. This particular class of WaMI kernels will be illustrated in the next section with a specific choice of nonlinear warpings involving cumulative distribution functions (CDFs) of beta distributions.

3.2. Examples. The flexibility of the WaMI-GP as a generative model is depicted in this section with various examples. In what follows we take for the base kernel k_β a radial kernel of the Matérn class with $\nu = 5/2$.

Stationary subcase. Here all univariate deformations are the identity. In Figure 6 we illustrate the warped space (here the overall warping amounts to A), the WaMI kernel, and corresponding GP sample paths with

$$A = \begin{pmatrix} 5 & 10 \\ 7.5 & 2.5 \end{pmatrix}. \tag{10}$$

Note that in case of an isotropic base kernel k_β, the first eigenvectors of $AA^\top$, ordered with increasing eigenvalues, give the directions of high variations appearing in the sample paths. This simple property could be used in a step-by-step parameter estimation procedure for choosing directions in which warping should be employed.

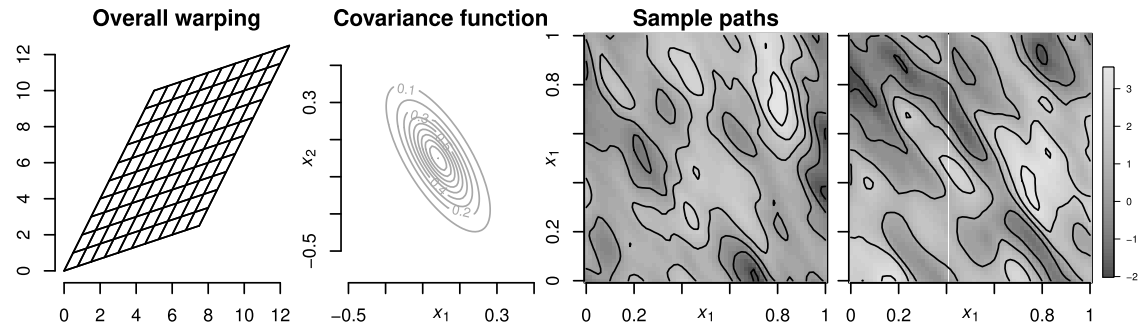

Figure 6. *WaMI-GP model in the case of a stationary base kernel and no axial deformation. From left to right: warping represented by mapping of the grid* $(\frac{i}{10}, \frac{j}{10})_{0\leq i,j\leq 10}$*, the covariance function* $c(\cdot,(0,0)^\top)$*, and two corresponding WaMI-GP realizations.*

Axial warping subcase. Before composing T with A, we now illustrate the case of axial deformations alone. We take A equal to the identity matrix. We keep T_2 as the identity function but $T_1 = I(\cdot;5,5)$, where $I(\cdot;\delta_1,\delta_2)$ is a CDF of a beta distribution. When it is not U-shaped, the density of the beta distribution is unimodal. We restrict here to this class of warpings in order to limit the number of parameters. For large designs, it is possible to increase the model flexibility, by considering more complicated parametrized warpings (as we illustrate later with a warping for modeling two high-variation regions) or using basis functions for warping estimation (as in, e.g., Xiong et al. [47]). Nonetheless, the CDF of a beta distribution provides a relatively diverse family of nonlinear deformations of a segment with only two shape parameters, δ_1, δ_2. This function also has been used in other situations for defining univariate deformations, e.g., in [41].

Moreover the warping may produce a very strong contraction of $[0,1]$ at the endpoints, as its derivative can be zero at 0 and 1 (for $\tau_{i,1}$, $\tau_{i,2} > 1$). To avoid such singularity, and relax this strong assumption, $I(\cdot;\tau_{i,1},\tau_{i,2})$ is combined with a linear function as

$$T_i(\cdot;\boldsymbol{\tau}_i) : x \to \frac{1}{1+\tau_{i,3}}\left(\tau_{i,3}x + I_{\tau_{i,1},\tau_{i,2}}(x)\right) \tag{11}$$

with $\tau_{i,3} \geq 0$. This parameter is empirically set to 1 here. In Figure 7, we observe that this covariance setting allows high variations in the vertical direction, at $x_2 = 1/2$ where the density of the axial warping is the highest.

Example of noncanonical orientation and two high-variation zones. Having a neutral parametrization toward canonical axes is the key idea for estimating arbitrary directions of heterogeneous variations. We now take

$$A = \begin{pmatrix} \cos(\pi/12) & -\sin(\pi/12) \\ \sin(\pi/12) & \cos(\pi/12) \end{pmatrix} \text{ and } T_2 = x + I(2x;15,15) + I(2(x-1/2);15,15), \tag{12}$$

which creates two variation zones. One can observe in Figure 8 the links between high-variation regions of realizations and the overall warping appearing in the corresponding WaMI kernel.

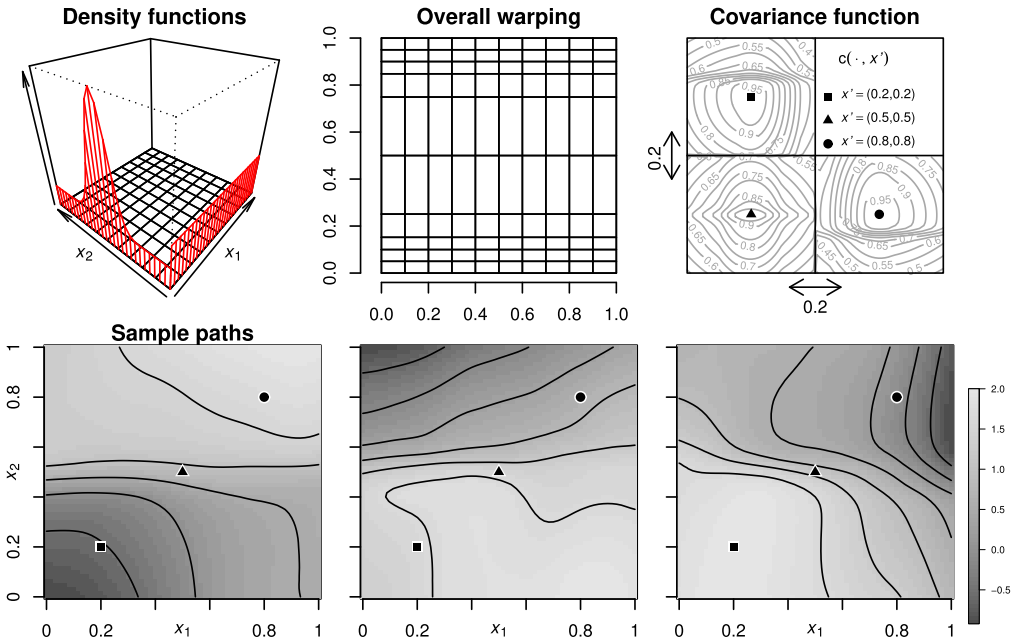

Figure 7. *WaMI-GP with axial deformations. From left to right: density function of the deformations in each direction, warping of the grid* $(\frac{i}{10}, \frac{j}{10})_{0\leq i,j\leq 10}$*, the covariance function* $c(\cdot, \boldsymbol{x}'))$ *for different values of* $\boldsymbol{x}'$*, and three corresponding WaMI GP realizations.*

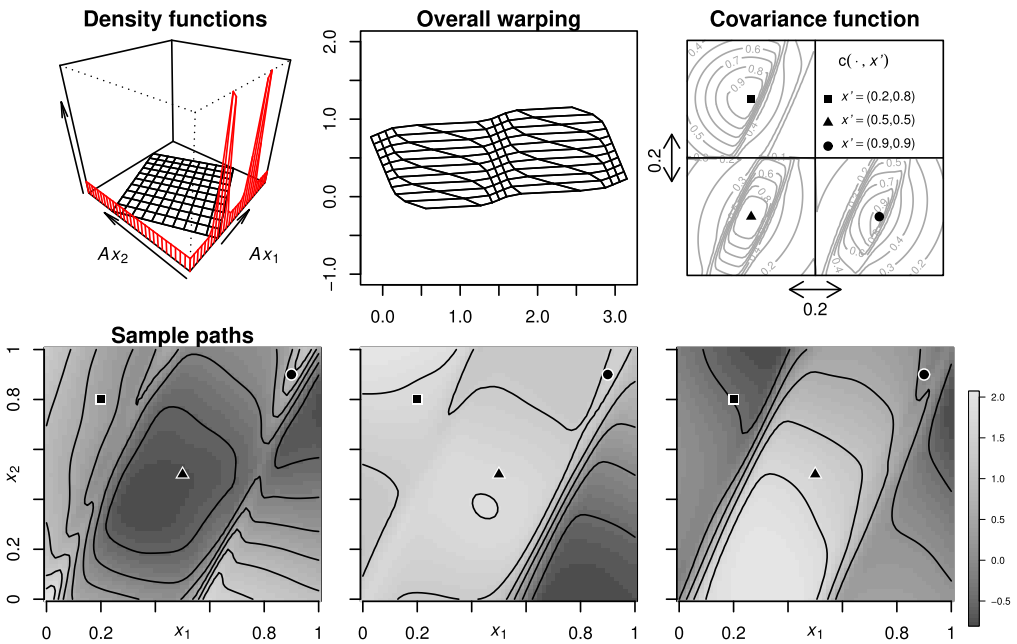

Figure 8. *Example of a WaMI-GP with two regions of high variations. From left to right: density functions of the deformations in each direction after the linear transformation, warping of the grid* $(\frac{i}{10}, \frac{j}{10})_{0\leq i,j\leq 10}$*, the covariance function* $c(\cdot, \boldsymbol{x}'))$ *for different values of* $\boldsymbol{x}'$*, and three corresponding WaMI-GP realizations.*

3.3. Strict positive-definiteness and differentiability properties. We prove here some properties of the WaMI kernel and the associated (centered) WaMI-GP. First we ensure that under conditions on k_β, A, and T_i's, the WaMI kernel is strictly positive definite. Although the strict definiteness is not necessary for a covariance function, this property is useful for avoiding singularity issues with covariance matrices.

Proposition 3.2 (strict positive definiteness). *Assume that k_β is strictly positive definite, that the $T_i(\cdot;\boldsymbol{\tau}_i)$ are injective, and that the rank of A is equal to d. Then the WaMI kernel of (7) is strictly positive definite.*

Proof. The injectivity of $T_i\ \forall\, i = 1,\dots,p$ implies the injectivity of T. Since A is injective, the composition of T with $\boldsymbol{x} \to A\boldsymbol{x}$ is injective, and the strict positive definiteness of $c_{\boldsymbol{\theta}}$ results from the same property of k_β. ∎

Let us now focus on differentiability questions. We give conditions for getting mean-squared differentiability and sample path differentiability of the WaMI-GP. We say that Z is mean-squared differentiable at a point $\boldsymbol{x} \in D$ in the ith canonical direction if there is a random variable $Z_i^{(1)}$ of order 2 ($\in L^2$) such that

$$\lim_{h\to 0}\left[\mathbb{E}\left(\left(\frac{Z_{\boldsymbol{x}+h\boldsymbol{e}_i} - Z_{\boldsymbol{x}}}{h} - Z_i^{(1)}\right)^2\right)\right] = 0. \tag{13}$$

The random vector $\nabla Z_{\boldsymbol{x}} = (Z_i^{(1)},\dots,Z_d^{(1)})^\top$, the gradient of Z at $\boldsymbol{x}$, will be used later in section 4 for the definition of new criteria.

Proposition 3.3 (mean-squared differentiability). *The centered GP with covariance c defined in (7) is mean-squared differentiable (i.e., is mean-squared differentiable in any direction) as soon as*

1. $\forall\, i \in \{1,\dots,q\}$, $T_i(\cdot;\boldsymbol{\tau}_i)$ *have regularity C^1 on $\mathbb{R}$,*
2. $\forall\, j,j' \in \{1,\dots,q\}$ *and* $\boldsymbol{u} \in \mathbb{R}^q$, $\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}|_{(\boldsymbol{u},\boldsymbol{u})}$ *exists and is finite.*

Proof. The warping T, whose components are the $T_i(\cdot;\boldsymbol{\tau}_i)$ functions, is C^1 on $\mathbb{R}^d$. Using the regularity of k_β and T, the chain rule applied to (7) gives that $\forall \boldsymbol{x} \in D$, $\frac{\partial^2 c(\boldsymbol{u},\boldsymbol{u}')}{\partial u_i \partial u'_i}|_{(\boldsymbol{x},\boldsymbol{x})}$ exists and is finite. Thus the corresponding GP is mean-squared differentiable (see, e.g., [24, p. 49]). ∎

Another relevant property when defining a covariance function is the almost sure differentiability of sample paths of the associated GP. In general finite-dimensional distributions of a stochastic process do not determine sample paths, and studying sample path properties from a covariance function calls for additional assumptions such as separability, as assumed here. In more generality, existence of separable versions is discussed in [8]; see, e.g., [24] for a summary.

Proposition 3.4 (sample path differentiability). *With the same assumptions on $T_i(\cdot;\boldsymbol{\tau}_i)$'s and k_β as in Proposition 3.3, and assuming in addition that D is compact and that there exist $C_0,\eta_0,\varepsilon_0 > 0$ such that $\forall j,j' \in \{1,\dots,q\}$, and $\forall \boldsymbol{u},\boldsymbol{u}' \in \mathbb{R}^q$, $||\boldsymbol{u}-\boldsymbol{u}'|| < \varepsilon_0$, we have $\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}|_{(\boldsymbol{u},\boldsymbol{u})} + \frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}|_{(\boldsymbol{u}',\boldsymbol{u}')} - 2\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}|_{(\boldsymbol{u},\boldsymbol{u}')} \le \frac{C_0}{|\ln||\boldsymbol{u}-\boldsymbol{u}'||^{1+\eta_0}}$. Then the covariance c provides a centered GP possessing a version with differentiable sample paths.*

Proof. We denote by T_i' the derivative of T_i. Let us take C, $\eta > 0$, and $0 < \varepsilon \leq 1/C_T$ (with C_T a Lipschitz constant of $\boldsymbol{x} \to T(A\boldsymbol{x})$) such that

1. $C = C_0 \sum_{j=1}^{q} \sum_{j'=1}^{q} a_{j1} a_{j'1} \sup_{\boldsymbol{x} \in D}(T_j'(\boldsymbol{a}_1^\top \boldsymbol{x})) \sup_{\boldsymbol{x} \in D}(T_{j'}'(\boldsymbol{a}_1^\top \boldsymbol{x}'))$,
2. $\forall \boldsymbol{x}, \boldsymbol{x}' \in D$, $||\boldsymbol{x} - \boldsymbol{x}'|| < \varepsilon$ implies $||T(A\boldsymbol{x}) - T(A\boldsymbol{x}')|| < \varepsilon_0$ (by continuity of T),
3. $\forall \boldsymbol{x}, \boldsymbol{x}' \in D$, $||\boldsymbol{x} - \boldsymbol{x}'|| < \varepsilon$ implies $\frac{1}{|\ln(C_T||\boldsymbol{x}-\boldsymbol{x}'||)|^{1+\eta_0}} \leq \frac{1}{|\ln||\boldsymbol{x}-\boldsymbol{x}'||^{1+\eta}}$ (by existence of the limit $\lim_{h\to 0}(\frac{\ln|\ln|h||}{\ln|\ln(C_T)+\ln|h||}(1+\eta_0) - 1) = \eta_0 > 0$).

Then we have $\forall\, \boldsymbol{x}, \boldsymbol{x}' \in D$, $||\boldsymbol{x} - \boldsymbol{x}'|| < \varepsilon$,

$$
\begin{aligned}
(14) \qquad & \left.\frac{\partial^2 c(\boldsymbol{u}, \boldsymbol{u}')}{\partial u_1 \partial u_1'}\right|_{(\boldsymbol{x},\boldsymbol{x})} + \left.\frac{\partial^2 c(\boldsymbol{u}, \boldsymbol{u}')}{\partial u_1 \partial u_1'}\right|_{(\boldsymbol{x}',\boldsymbol{x}')} - 2 \left.\frac{\partial^2 c(\boldsymbol{u}, \boldsymbol{u}')}{\partial u_1 \partial u_1'}\right|_{(\boldsymbol{x},\boldsymbol{x}')} \\
& = \sum_{j=1}^{q} \sum_{j'=1}^{q} a_{j1} a_{j'1} T_j'\left(\boldsymbol{a}_1^\top \boldsymbol{x}\right) T_{j'}'\left(\boldsymbol{a}_1^\top \boldsymbol{x}'\right) \left(\left.\frac{\partial^2 k_\beta(\boldsymbol{v}, \boldsymbol{v}')}{\partial v_j \partial v_{j'}'}\right|_{\substack{(T(A\boldsymbol{x}'),\\ T(A\boldsymbol{x}'))}} \right. \\
& \qquad \left. + \left.\frac{\partial^2 k_\beta(\boldsymbol{v}, \boldsymbol{v}')}{\partial v_j \partial v_{j'}'}\right|_{\substack{(T(A\boldsymbol{x}),\\ T(A\boldsymbol{x}))}} - 2 \left.\frac{\partial^2 k_\beta(\boldsymbol{v}, \boldsymbol{v}')}{\partial v_j \partial v_{j'}'}\right|_{\substack{(T(A\boldsymbol{x}),\\ T(A\boldsymbol{x}'))}} \right) \\
& \leq \frac{C}{|\ln||T(A\boldsymbol{x}) - T(A\boldsymbol{x}')||^{1+\eta_0}} \leq \frac{C}{|\ln||\boldsymbol{x} - \boldsymbol{x}'||^{1+\eta}}.
\end{aligned}
$$

Using the theorem of sample path continuity for GP derivatives (see, e.g., [40, p. 55], or the supplementary material with the notations of the article), we get the sample path continuity for the GP $\partial Y/\partial x_1$ and thus ∇Y by generalizing to all components. ∎

Remark 1. These properties can be extended to a higher order of differentiation with equivalent hypotheses on a higher order of differentiability for the $T_i(\cdot; \boldsymbol{\tau}_i)$'s and k_β.

Remark 2. In this article, the estimation of A and of the $\boldsymbol{\tau}_i$ parameters ($i = 1, \ldots, p$) are performed by maximum likelihood; gradients are calculated analytically and the numerical optimization relies on the BFGS algorithm with one or several initial points, using the R package *kergp* [7].

3.4. WaMI-GP interpolation and sequential design on the running example. Let us now come back to our running example function (3). For the sake of brevity, we directly look at the results of an MSE-driven sequential design of experiments under the WaMI-GP model compared to three competing models covered in the last section: stationary GP, GP with axial warping, and TGP. The WaMI covariance is parametrized with the T_i's as in (11) and with A and k_β as in (9). In Figure 9, we display the absolute difference between the real function and predictions from the four models after 20 sequential design steps based on the MSE criterion. The sequential sampling procedures start with models built on an initial design of size 8 (an LHS maximized with maximin distance). Looking at the selected points along the four competing sequential designs, we see that the MSE design relying on the WaMI-GP model allocates more evaluations in the high-variation region (around the line of equation $0.75 = x_1 + x_2$) and fewer evaluations in the flat regions (upper right). For the other models, prediction errors tend to occur in the high-variation region.

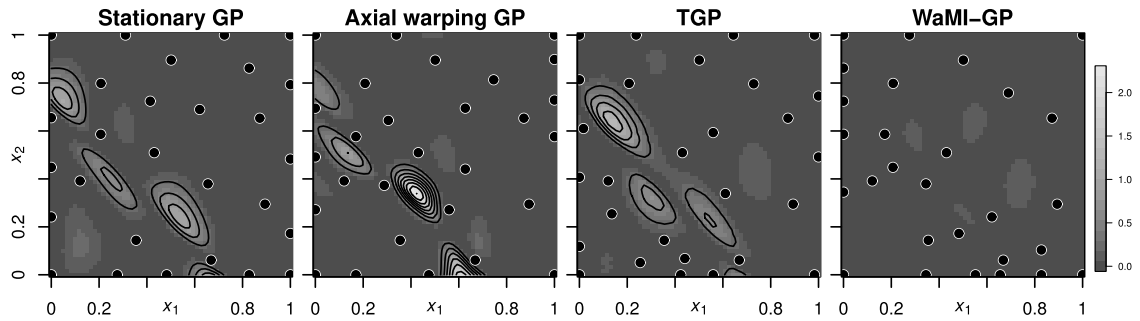

Figure 9. *Prediction errors of four competing models on the running example function. The different models are a stationary anisotropic GP, an axial warping GP, TGP, and WaMI-GP. For each method, we see the* 20*th step of a sequential design driven by the MSE criterion (shared initial design).*

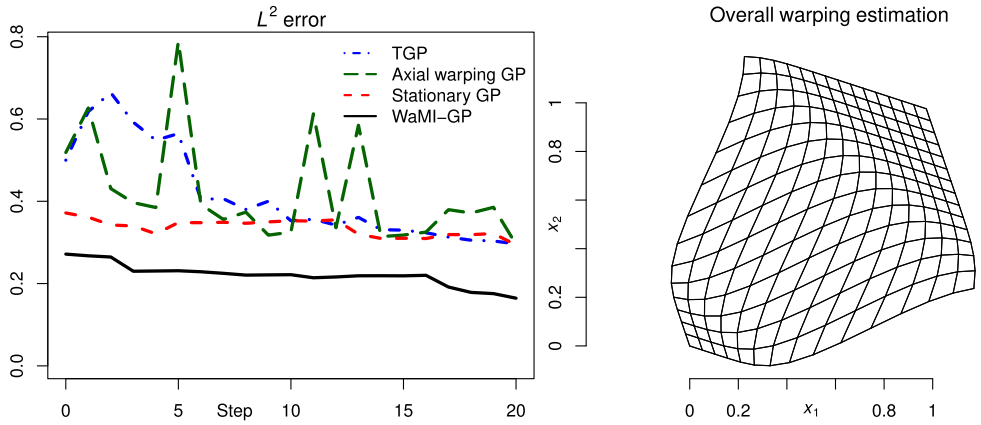

Figure 10. *Prediction error of the running example function by the four considered models at each step of MSE-driven sequential designs. Estimated overall warping (i.e.,* $\boldsymbol{x} \to T(A\boldsymbol{x})$*) extracted from the WaMI model.*

Hence our model, by detecting the high-variation region and also associating a higher MSE there, enables us to comparatively achieve enhanced prediction performance as illustrated in Figure 10.

We investigate how the choice of the WaMI covariance function influences the uncertainty on the predictions. From the synthetic example, we consider the coverage probability over 50 LHS designs of size 20. The 95% confident intervals provided by the stationary model on a 40×40 grid contain the true value for 65.6% of the points (in average over the point location and the 50 designs of size 20), while the WaMI model is less permissive with a ratio of 90.0%.

These first results illustrate that WaMI-GP is able to account for heterogeneous regions in a semiautomated way (here all parameters including axes are estimated by MLE but the base kernel and the number of warping dimensions is fixed in advance). Two applications, presented in subsection 5.1 and 5.2, will be used to complement these results in subsection 5.3.

Here the performance of variance-based sequential criteria is improved with WaMI-GP provided that some prior knowledge is available regarding the heterogeneities of the unknown

function. In contrast, available information on the objective function might be limited. For this reason we explore in the next section an approach where the prior covariance is arbitrary and the emphasis is put instead on sampling criteria. Later on in subsection 5.4 the two approaches of working on kernels or on the sampling criteria for learning functions with heterogeneous variations will be combined within a numerical benchmark.

4. Novel sampling criteria for detection of high variations. The example in the previous section showed that exploring high-variation regions is key to quickly reducing the overall approximation error. We explore here classes of strategies to do so, where the prior covariance is arbitrary (it may be a stationary one, a WaMI, or any other kind of kernel) and the targeted exploration is driven by specifically designed sampling criteria.

The goal of intensifying exploration in high-variation regions encourages us to invest more credit in regions where the posterior distribution indicates more local variability (or where sampling could reduce relevant measures of variability). A limitation with usual variance-based criteria, however, is that they are homoscedastic in the observations, i.e., they depend solely on the geometry of the experimental design and not on the response values. Hence trying to locate high-variation regions with variance-based criteria does not make much sense, unless the model accounts for heterogeneities through estimated parameters that reflect them, such as with WaMI-GP. Our approach here, assuming that the GP possesses sufficient differentiability properties, is to rely instead on the gradient of the GP in order to add points in unexplored regions with potentially high slopes.

The starting point is to acknowledge that, under sufficient regularity conditions, $(\nabla Y_{\boldsymbol{x}})_{\boldsymbol{x}\in D}$ is a vector-valued GP and that its conditional distribution knowing $\mathcal{A}_n$ is driven by derivatives of m_n and c_n (see, e.g., Theorem 5.3.10 of [40]). Assuming indeed the differentiability of m_n and the existence $\forall\, i$ of derivatives $\frac{\partial^2}{\partial t_i \partial t'_i} c_n(\boldsymbol{t},\boldsymbol{t}')|_{\boldsymbol{t}=\boldsymbol{t}'=\boldsymbol{x}}\ \forall\ \boldsymbol{x},\boldsymbol{x}' \in D$,

$$\mathbb{E}\left(\nabla Y_{\boldsymbol{x}} \mid \mathcal{A}_n\right) = \nabla m_n(\boldsymbol{x}),\ \text{and}\ \operatorname{cov}\left(\nabla Y_{\boldsymbol{x}} \middle| \mathcal{A}_n\right) = \left(\frac{\partial^2}{\partial t_i \partial t'_j} c_n(\boldsymbol{t},\boldsymbol{t}') \middle|_{\boldsymbol{t}=\boldsymbol{x},\boldsymbol{t}'=\boldsymbol{x}'} \right)_{1\le i,j\le d}. \tag{15}$$

By exploiting the distribution in (15), several scalar indicators can be defined to quantify local variations and related uncertainties. In this work we chose to focus essentially on variance-based criteria for (exponentiated) gradient norms. Therefore, we consider the squared gradient norm process $(Q_{\boldsymbol{x}})_{\boldsymbol{x}\in D}$ defined by

$$Q_{\boldsymbol{x}} = ||\nabla Y_{\boldsymbol{x}}||^2_{\mathbb{R}^d} = \nabla Y_{\boldsymbol{x}}^\top \nabla Y_{\boldsymbol{x}}. \tag{16}$$

Although the squared gradient norm is obtained by applying a simple operation (taking the squared Euclidean norm) to a vector-valued GP, working out its distribution is not straightforward. Actually, even by fixing $\boldsymbol{x}$, working out the probability distribution of quadratic forms in arbitrary Gaussian variables is involved and while it is tempting to build on such a distribution for sequential design, coming up with tractable sampling criteria is more demanding than in the Gaussian case. Yet, as we develop next, some (fractional) moments of $Q_{\boldsymbol{x}}$ can be calculated in closed form or computed efficiently, leading to practical infill sampling criteria. Let us generalize indeed MSE and IMSE criteria to the (exponentiated) gradient norm.

Definition 4.1 (gradient norm variance criterion and generalizations). *Given n function evaluation results and $\boldsymbol{x} \in D$, we define the gradient norm variance (GNV) criterion as*

$$J_n^{\mathrm{GNV},\eta}(\boldsymbol{x}) = \mathrm{var}\left(||\nabla Y_{\boldsymbol{x}}||^{\eta} \mid \mathcal{A}_n\right) = \mathrm{var}\left(Q_{\boldsymbol{x}}^{\eta/2} \,\middle|\, \mathcal{A}_n\right), \tag{17}$$

where the norm is elevated to some power $\eta > 0$. Note that while the transformed norm loses its homogeneity, we abusively refer to this criterion as "GNV with exponent η" or "$\mathrm{GNV}(\eta)$*."* *Besides this, this class of criteria can also be generalized in the same way as* IMSE *generalizes* MSE *by integration, defining* IGNV *by*

$$J_n^{\mathrm{IGNV},\eta}(\boldsymbol{x}) = \int_{\boldsymbol{u}\in D} \mathbb{E}\left(\mathrm{var}\left(Q_{\boldsymbol{u}}^{\eta/2} \,\middle|\, \mathcal{A}_n, Y_{\boldsymbol{x}}\right) \middle| \mathcal{A}_n\right) \mathbf{d}\boldsymbol{u}. \tag{18}$$

The following property gives a close formula for GNV in the case $\eta = 2$ and semianalytical in the case $\eta = 1$, followed by integral formulas for the corresponding IGNV criteria.

Proposition 4.2. *Let $\boldsymbol{x} \in D$ and denote by $(\lambda_i(\boldsymbol{x}))_{1\leq i\leq d}$ the eigenvalues of $\nabla\otimes\nabla^{\top} c_n(\boldsymbol{x},\boldsymbol{x})$. Then, the the* $\mathrm{GNV}(2)$ *and* $\mathrm{GNV}(1)$ *criteria can be written as follows:*

$$J_n^{\mathrm{GNV},\eta=2}(\boldsymbol{x}) = 4\ \nabla m_n(\boldsymbol{x})^{\top} \nabla\otimes\nabla^{\top} c_n(\boldsymbol{x},\boldsymbol{x}) \nabla m_n(\boldsymbol{x}) + 2\sum_{i=1}^{d} \lambda_i(\boldsymbol{x})^2, \tag{19}$$

$$J_n^{\mathrm{GNV},\eta=1}(\boldsymbol{x}) = ||\nabla m_n(\boldsymbol{x})||^2 + \mathrm{tr}\left(\nabla\otimes\nabla^{\top} c_n(\boldsymbol{x},\boldsymbol{x})\right) - \mathbb{E}\left(\sqrt{Q_{\boldsymbol{x}}}|\mathcal{A}_n\right)^2. \tag{20}$$

The corresponding integral criteria can be expanded as

$$\begin{aligned} J_n^{\mathrm{IGNV},\eta=1}(\boldsymbol{x}) = &\int_D \left(||\nabla m_n(\boldsymbol{u})||^2 + \frac{1}{c_n(\boldsymbol{x},\boldsymbol{x})} \kappa_n(\boldsymbol{u},\boldsymbol{x})^{\top} \kappa_n(\boldsymbol{u},\boldsymbol{x})\right) \mathbf{d}\boldsymbol{u} \\ &+ \int_D \left(\mathrm{tr}\left(\nabla\otimes\nabla^{\top} c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u})\right) - \mathbb{E}\left(\mathbb{E}\left(\sqrt{Q_{\boldsymbol{u}}}|\mathcal{A}_n, Y_{\boldsymbol{x}}\right)^2 \middle| \mathcal{A}_n\right)\right) \mathbf{d}\boldsymbol{u}, \end{aligned} \tag{21}$$

$$\begin{aligned} J_n^{\mathrm{IGNV},\eta=2}(\boldsymbol{x}) = &\int_{\boldsymbol{u}\in D} \left(4\nabla m_n(\boldsymbol{u})^{\top} \nabla\otimes\nabla^{\top} c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \nabla m_n(\boldsymbol{u}) + 2\sum_{i=1}^{d} \lambda_{i,\boldsymbol{x}}(\boldsymbol{u})^2\right) \mathbf{d}\boldsymbol{u} \\ &+ \frac{4}{\mathrm{var}\left(Y_{\boldsymbol{x}}|\,\mathcal{A}_n\right)} \int_{\boldsymbol{u}\in D} \kappa_n(\boldsymbol{u},\boldsymbol{x})^{\top} \nabla\otimes\nabla^{\top} c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \kappa_n(\boldsymbol{u},\boldsymbol{x})\ \mathbf{d}\boldsymbol{u}, \end{aligned} \tag{22}$$

where $\lambda_{i,\boldsymbol{x}}(\boldsymbol{u})$ are the eigenvalues of $\nabla\otimes\nabla^{\top} c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) = \mathrm{cov}(\nabla Y_{\boldsymbol{u}}|\mathcal{A}_n, Y_{\boldsymbol{x}})$ and $\kappa_n(\boldsymbol{u},\boldsymbol{x})$ is the vector of covariances between the components of $\nabla Y_{\boldsymbol{u}}$ and $Y_{\boldsymbol{x}}$ knowing $\mathcal{A}_n$.

A proof is provided in Appendix B.

Remark 3. For $\boldsymbol{u}$, $\boldsymbol{x}$ in D, the expectation terms are approximated by quadrature formulas of univariate or bivariate integrals:

$$\mathbb{E}\left(\sqrt{Q_{\boldsymbol{x}}}\right) = \int_{\mathbb{R}} \sqrt{q} f_Q\left(q; \nabla m_n(\boldsymbol{x}), \nabla\otimes\nabla^{\top} c_n(\boldsymbol{x},\boldsymbol{x})\right) \mathrm{d}q, \tag{23}$$

$$\mathbb{E}\left(\mathbb{E}\left(\sqrt{Q_{\boldsymbol{u}}}|\mathcal{A}_n, Y_{\boldsymbol{x}}\right)^2 \middle| \mathcal{A}_n\right) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \sqrt{q} f_Q\left(q; \boldsymbol{\mu}_n(y;\boldsymbol{u},\boldsymbol{x}), \Gamma_n(\boldsymbol{u},\boldsymbol{x})\right) \mathrm{d}q\right)^2 \varphi_n(y;\boldsymbol{x}) \mathrm{d}y,$$

with $\boldsymbol{\mu}_n(y;\boldsymbol{u},\boldsymbol{x}) = \nabla m_n(\boldsymbol{u}) + \frac{y-m_n(\boldsymbol{x})}{c_n(\boldsymbol{x},\boldsymbol{x})}\kappa_n(\boldsymbol{u},\boldsymbol{x})$, $\Gamma_n(\boldsymbol{u},\boldsymbol{x}) = \nabla\otimes\nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u})$, and $\varphi_n(\cdot;\boldsymbol{x})$ the normal probability density function of $Y_{\boldsymbol{x}}$ (mean $m_n(\boldsymbol{x})$ and variance $c_n(\boldsymbol{x},\boldsymbol{x})$). We used the R package package *CompQuadForm* [9] for computing the distribution $f_Q(\cdot;\boldsymbol{\mu},\Gamma)$ of the quadratic form $Q = \boldsymbol{Z}^\top\boldsymbol{Z}$, $\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{\mu},\Gamma)$.

Remark 4. When focusing on the variance of the gradient, a criterion may unfortunately evaluate preferably in regions with low amplitude, if these regions have high variations. As the variance of the gradient is proportional to the variance of the GP, this effect may be limited, but it seems important to be aware of it when using derivative-based criteria.

Figure 11 displays the values of the four criteria, GNV and IGNV for $\eta = 1, 2$, for the running example test function (3) with the GP model of Figure 2. Integrated criteria (IMSE, IGNV) require more computational resources but can be preferred for their generally smoother variations, and also lower values at the edges of the input space compared to MSE and GNV. As expected from variance-based criteria, they do not provide a higher criterion value for the high-variation region in the bottom left quarter of the input space. On the contrary, gradient-based criteria provide higher values where f has high variations. In case of integrated gradient-based criteria, surroundings of evaluation points are penalized. These observations suggest that integrated gradient-based criteria should be useful in order to perform some kind of compromise between global uncertainty reduction and a focus on high variations. These criteria will be tested with two applications in subsection 5.4, to illustrate their ability to detect heterogeneous regions.

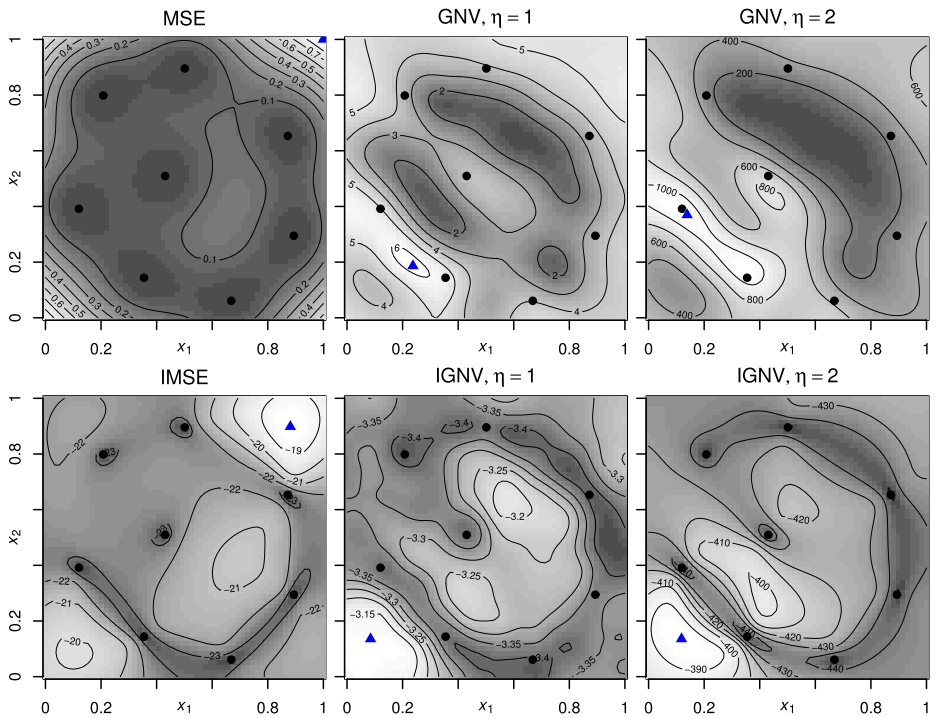

Figure 11. *Classical and proposed criteria according to a stationary GP modeling.*

5. Applications. We consider two applications coming from an IRSN case study on nuclear safety and from a NASA case study on fluid mechanics. A special attention is devoted to the assessment of the capability of the methodological contributions developed in this paper for the approximation of functions with heterogeneous variations and to their comparison with existing approaches, either in terms of criteria (MSE, IMSE, $\text{GNV}_{\eta=1,2}$, $\text{IGNV}_{\eta=1,2}$) or in terms of surrogate models (stationary anisotropic, TGP, WaMI-GP). This assessment is achieved by focusing on estimates of the L^2 prediction error: $\Delta = (\int_D (\mu(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x})^{1/2}$, with μ one or the other predictor based on some experimental design strategy. Experimental design strategies are replicated by starting from different initial designs, as detailed next.

We also consider a simplification of the $\text{IGNV}_{\eta=1,2}$ criteria, where the mean value of $Y_{\boldsymbol{x}}$ is plugged into the integrand, more precisely,

$$J_n^{\text{IGNV}_{\text{plugin}},\eta}(\boldsymbol{x}) = \int_{\boldsymbol{u}\in D} \text{var}\left(||\nabla Y_{\boldsymbol{u}}||^{\eta} \,|\, \mathcal{A}_n, Y_{\boldsymbol{x}} = m_n(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{u}. \tag{24}$$

5.1. Cracking simulation of heterogeneous materials. This test case concerns mechanical studies in nuclear installations. Their objective is to analyze the crack propagation inside a heterogeneous material such as concrete using the IRSN *Xper* code [28]. Two input variables, related to geometrical and mechanical properties of the material, are considered here: the ratio of interface energy W and the inclusion length L. The output of interest is the cracking energy, i.e., the smallest energy required to break the material apart. Simulation times are long, from one day to one week, and therefore evaluations should be chosen carefully in order to capture the function behavior. The available dataset includes 216 points corresponding to the simulation of the response on a 36×6 grid (Figure 12). We observe a region where a small variation of the inputs impacts drastically the output. This high-variation zone is located along a straight line, slightly nonaligned to the canonical axes. Several series of tests

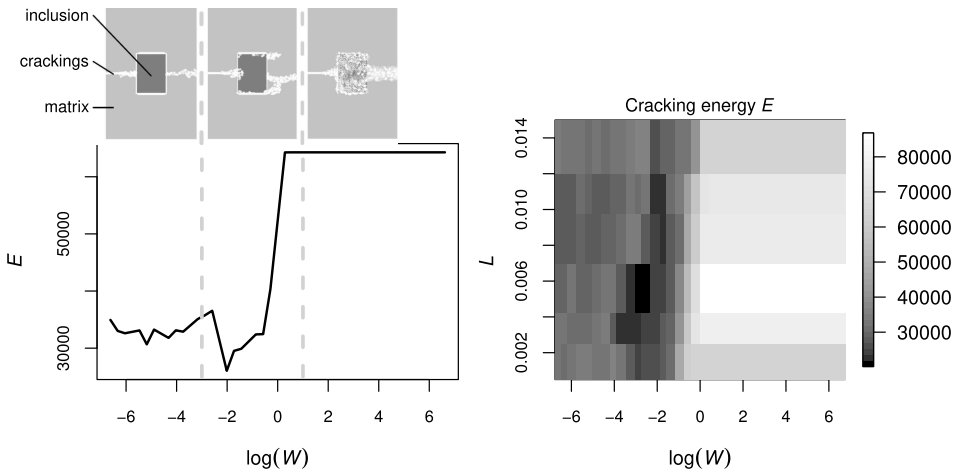

Figure 12. *Cracking energy of a heterogeneous material depending on mechanical parameters. The images (top left) represent three different crackings of a component. We see that it propagates around (left) or through (right) the inclusion depending on the input. These two modes correspond respectively to low or high cracking energy. A transition zone appears in between with high variations.*

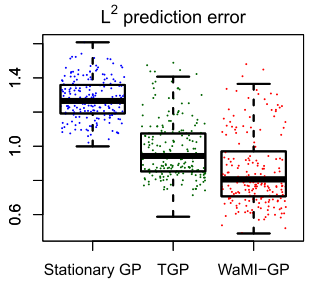

Figure 13. *Comparison of L^2 prediction errors on the IRSN test case between the three candidate models: stationary anisotropic GP, TGP, and WaMI-GP. The boxplots are obtained from repetitions with* 5000 *different initial designs (more detail in the text).*

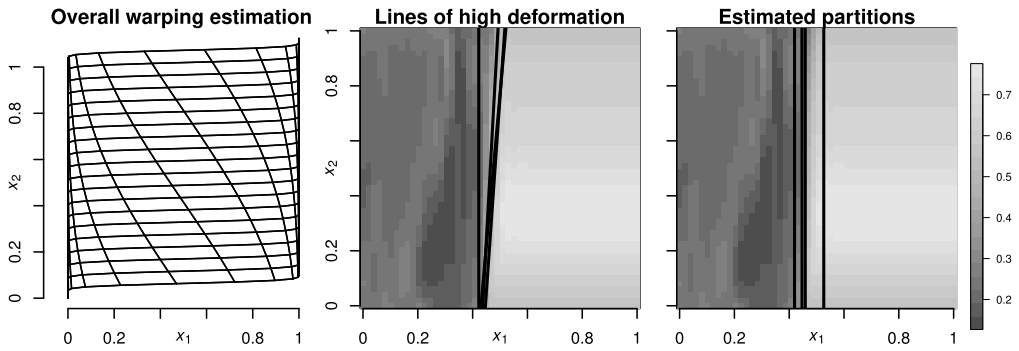

Figure 14. *Some features of models with median prediction errors. Left: estimated warping of the WaMI-GP model; middle: lines of maximal distortion for five (most) median models; right: lines of partitioning for five median TGP models.*

are conducted. For the sake of simplicity, the input variables are rescaled between 0 and 1 and are denoted by x_1 and x_2.

We first compare the predictive performances of stationary GP, WaMI-GP, and TGP methods. The WaMI covariance is parametrized with the T_i's as in (11) and with A and k_β as in (9). We built 5000 space filling designs of size 20 (optimized with a maximin criterion; see, e.g., [10]). For each initial design, predictions are performed with the three competing models, in a noise-free setting. Figure 13 shows that our approach outperforms the other two in terms of L^2 prediction errors.

The estimated (overall) warpings are displayed in Figure 14 (we take the warping from the design giving a median prediction errors). It appears that, as expected, our model dilates the space around the high-variation region. We also display in the input space, the lines of maximal distortion (where the determinant of the Jacobian matrix of the warping is maximal), and the lines partitioning the input space in the TGP method. These lines are both in the same area, meaning that both methods can detect the high-variation region. However, since the method allows linear transformation of the input space, they are not exactly vertical in the case of the WaMI-GP, adapting with more freedom their directions to the shape of the actual high-variation region.

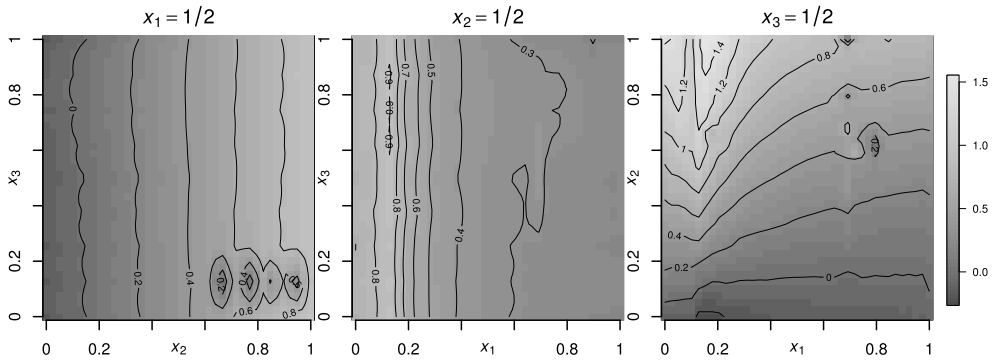

Figure 15. *Simple three-dimensional interpolation of the available data on the Langley Glide-Back Booster simulation test case (only three slices of the input cube are displayed).*

5.2. The Langley Glide-Back Booster simulation. The Langley Glide-Back Booster is a rocket booster developed at NASA. Its behavior is studied via numerical simulations. More details on the system behavior and purpose are provided in [34]. Three input variables (rescaled between 0 and 1) are considered: mach number (x_1), angle of attack (x_2), and sidelip angle (x_3). The output of interest is the lift force. An interpolation of the available data is displayed in Figure 15. We see that variations are mainly directed by canonical axes. A high-variation region is concentrated around the plane of equation $x_1 = 0.1$ (i.e., around mach 1). This calls for a nonstationary model. All models are considered in noisy settings in order to smooth out the prediction errors, as some discontinuity is observed, for example, at the bottom right of the first plot (region $x_1 \approx 0.5, x_2 \geq 0.5, x_3 \leq 0.5$). This is due to the complexity of the simulator whose convergence depends on a solver which sometimes returns inaccurate values despite automatic checks [17].

On this test case, the WaMI covariance is parametrized with the T_i's as in (11) and k_β is as in (9). The rotation matrix A is fixed to the identity: allowing rotation was tested first but it does not improve the results as on this data set the heterogeneous variations are mainly aligned with canonical axes.

5.3. Benchmark with nonsequential designs. We now make a study of the three data sets (the synthetic function (3), the IRSN test case, and the NASA test case) to evaluate the WaMI-GP model separately from the sampling approaches.

Several tests are performed for nonsequential designs including from 5 to 80 points in two-dimensional test cases and from 50 to 700 points for the three-dimensional test case. To account for stochastic effects due to the choice of the design, all experiments are repeated 50 times with different optimized space-filling designs. We then focus on the 5%, 50%, and 95% quantiles of the errors. The results are displayed in Figure 16.

We observe that in most cases, the WaMi-GP model has the lowest prediction error. Focusing on the NASA test case, it turns out that for a small training dataset, WaMI-GP leads to similar predictive performance as TGP in terms of median error. When increasing the number of points in the initial design, TGP clearly outperforms WaMI-GP. This makes sense as TGP model increases its complexity (i.e., its number of partitions and estimated

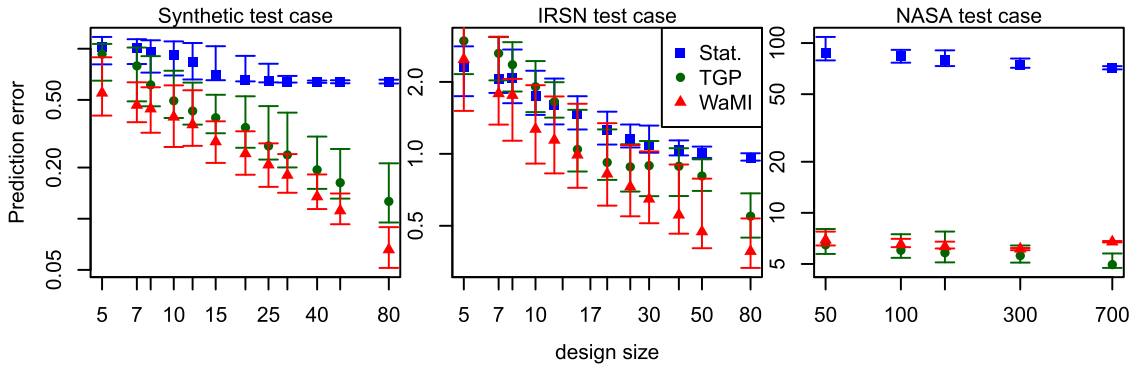

Figure 16. *Prediction error on the three datasets (one synthetic and two real) which randomizes over* 50 *LHS training sets of various sizes (*0.05*,* 0.5*, and* 0.95 *quantiles).*

parameters) according to the data while, in its present form, WaMI-GP has a fixed structure prescribed by the user.

On the computational effort, as we use the WaMI covariance within a standard form of GP modeling, the training cost has $O(n^3)$ complexity. TGP is an exception, where the division of the dataset leads generally to a much faster parameter estimation. Significant efforts have been made to reduce computing efforts for standard GP models (see, e.g., [31] and references therein), and this aspect was not a priority during our developments as we targeted applications with expensive-to-evaluate functions. For both WaMI-GP and TGP models, it takes a few seconds to estimate the model from 20 points and make 1600 predictions. However, for a design of about 100 points, estimation and predictions take about a couple of minutes for a WaMI-GP model (and about 10 times less under a TGP model). Work is in progress to optimize the current implementation of the WaMI-GP model.

5.4. Sequential designs for the IRSN and NASA test cases.

5.4.1. IRSN test case. For each criterion (MSE/IMSE, $\mathrm{GNV}_{\eta=1,2}$/$\mathrm{IGNV}_{\eta=1,2}$), we repeat 10 steps of the sequential design: point selections coupled with model updates. We choose to take as GP models a stationary isotropic and a WaMI one. The whole workflow is replicated 100 times with a space-filling initial design of $n = 20$ points uniformly drawn among optimized LHS designs. The results are displayed in Figure 17 and Table 1.

Let us first notice that the WaMI-GP model leads generally to the smallest prediction errors since it is adapted to the function f exhibiting a steep transition region.

When the model is stationary, the MSE and IMSE criteria do not focus on adding points in the steep transition region (one can say these methods explore D in a space-filling way). On the contrary, the $\mathrm{IGNV}_{\eta=1}$ criterion detects regions where the gradient's norm is high, leading to a better model training and to a reduction of 50% of the number of points (and therefore of simulations with the computer code) required to reach the same median error. When the model is nonstationary and well-adapted to the behavior of f, the IMSE focuses naturally on the high-variation zone and allows a reduction of about 30% in the number of simulations compared to the stationary framework. Finally, coupling WaMI-GP modeling and gradient-based criterion leads to rather poor results in Figure 17 since, by construction, both aspects contribute to exploitation with detrimental consequences on the exploration side.

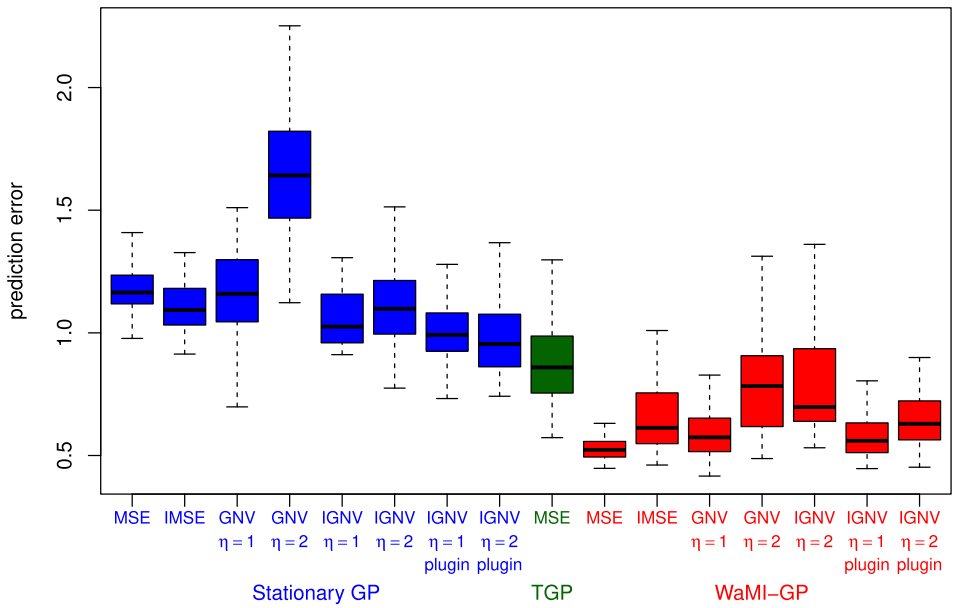

Figure 17. *Distribution of the prediction error after different sequential design of experiments. Sampling criteria are compared in both stationary and our nonstationary models.*

Table 1

Required number of steps for reaching a median error (computed from the 100 *initial designs) below a reference value of* 1.405 *(the value of the median error after six evaluations sampled with IMSE criterion and stationary model), with respect to the choice of model and criterion.*

	MSE	IMSE	GNV, $\eta = 1$	IGNV, $\eta = 1$ plugin	GNV, $\eta = 2$	IGNV, $\eta = 2$ plugin
Stationary GP model	10	6	>10	3	9	4
WaMI-GP model	5	4	>10	4	9	4

5.4.2. NASA test case. From initial designs of size 50, we perform 20 new evaluations chosen by MSE maximization. Results obtained in prediction with TGP and the WaMI-GP model are presented in Figure 18. We see that the prediction errors are reduced faster using our model. Indeed, the estimated warping allows us to dilate the input space in the region of high variations (around mach 1). It results in an increased model variance in this area and thus a denser exploration of it via MSE maximization. Note that the TGP method combined with the MSE criterion also leads to search patterns focusing in high-variation regions, as each partition has a GP with different variance levels (see, e.g., [18]). We also compare a gradient-based criterion, IGNV, $\eta = 2$, with a classical criterion IMSE relying on a stationary anisotropic GP model (Figure 18).

The IGNV($\eta = 2$) criterion leads to slightly lower prediction errors than the IMSE criterion based on the small budget of 20 points in dimension 3. Even if moderate, this improvement can be attributed to a more intense sampling of the high-variation region with the gradient-based

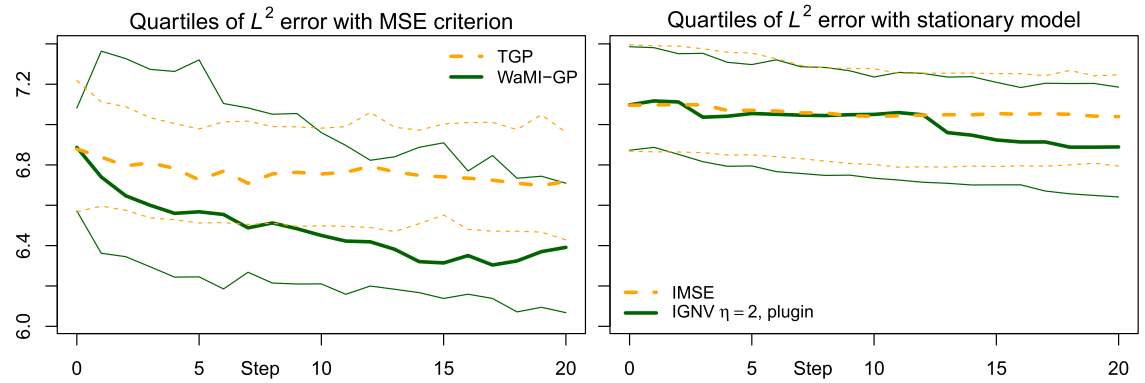

Figure 18. *Medians (bold line) and quartiles of prediction errors during sequential designs of experiments. Left: comparison of models TGP and WaMI-GP with a common criterion MSE. Right: comparison of criterion IMSE and IGNV, $\eta = 2$, with a standard stationary GP model.*

criterion. To conclude this experimental section, reinforcing exploration in high-variation regions appears to be a sound option to improve predictivity of surrogate models such as GPs, be it through adapted nonstationary covariances or via sampling criteria dedicated to this goal.

6. Conclusion. In this paper we introduced a nonstationary GP model and sampling approaches for prediction and design of experiments when the expensive-to-evaluate function presents heterogeneous variations. The proposed WaMI-GP (warped multiple index Gaussian process) model extends existing modeling approaches such as multiple index modeling and the nonlinear map method. We presented conditions under which the WaMI kernel is strictly positive definite and the corresponding centered GP is mean-squared differentiable or possesses differentiable sample paths almost surely. We applied the model on toy examples and on engineering case studies in dimensions 2 and 3. Although the number of parameters of WaMI-GP is kept affine (rather than exponential) with the dimension, the component-by-component univariate warpings lead to competitive performances with respect to stationary GP and TGP modeling. With larger data sets, we observed better performances of the TGP model in the second engineering test case. For smaller initial data sets, WaMI-GP and TGP obtained comparable performances at the start, but WaMI-GP proved better at approximating the response as more points were added by MSE maximization. It is also relevant to point out that in the case of a high-variation zone slightly not aligned with a canonical axis, our model is favored because its linear component can estimate an appropriate rotation of the data before the nonlinear warping (note that nonaxial partition in TGP is possible). In contrast, our method directly inherits from the nonlinear map method the ability to estimate an input space warping. This change of variables, dilating the space where there are high variations, and contracting smooth areas, can be used by practitioners as a tool for working out and visualizing "stationarization."

From a different viewpoint, we developed novel criteria in the sequential design of experiments for exploring functions with high-variation regions. These criteria are based on the GNV of the modeling GP. They are designed to sample preferably in high-variation regions,

where prediction errors are typically higher, but still performing a global exploration of the input space. We applied them to adaptively approximate functions arising from the two engineering case studies. When the covariance of the GP model is a priori stationary, some of the proposed criteria lead to a better prediction than MSE and IMSE thanks to their focus on steep regions. When combining the novel criteria with WaMI-GP, however, the effects are somehow cumulated and new evaluations are mostly concentrated around the high-variation region, leading to predictions that are less trustworthy when looking at performances over the whole domain.

This work paves the way to further research on sequential design of experiments for functions with heterogeneous variations, be it through the incorporation of nonstationarity within the models themselves, through targeted sampling criteria, or combinations of both. Perspectives include the definition of additional classes of criteria, relying, for instance, on higher GP derivatives, the stepwise uncertainty reduction paradigm [2], weighted IMSE approaches [29], or other. Also, batch-sequential versions of the proposed criteria and their extensions ought to be defined and worked out. Further work is needed to benchmark performances of the novel criteria in higher dimensions. Finally, the WaMi-GP model could be improved along several directions. This notably includes investigations into its relevance in higher dimensions as well as its estimation by efficient algorithms beyond brute force likelihood maximization. Notably, the estimation of the coefficients here is computed with likelihood maximization, while a full Bayesian approach might lead to more sensible results.

Appendix A. Nonlinear map with Gibbs' method. The warping to estimate is defined as $T_\tau(\boldsymbol{x}) = \boldsymbol{x}_0 + (\int_{P_{01}} g_i(\boldsymbol{u})\mathrm{d}s)_{1\le i\le d}^\top$, with P_{01} a predefined path between $\boldsymbol{x}_0$ and $\boldsymbol{x}$, for example, the corresponding segment. These density functions are expressed as linear combinations of radial basis functions. We see in Figure 19 how this method allows an approximation of a given deformation. We observe a degradation of the warping approximation with decreasing numbers of parameters to estimate: with a grid of 16 basis functions, i.e., with 32 parameters in dimension 2, the approximation fails. Here about 100 basis functions are needed to capture the nonstationarity in the whole domain.

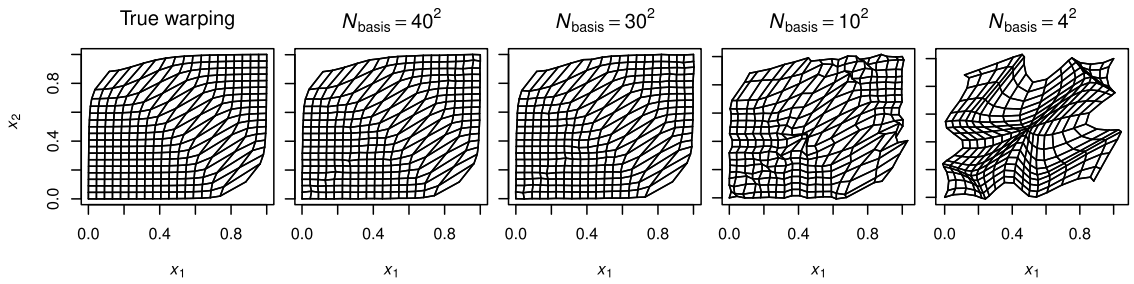

Figure 19. *Warping approximation with Gibbs' method, for different numbers of basis functions. An arbitrary warping ($T_0(\boldsymbol{x}) = \boldsymbol{x} + 1/10 \arctan(30(x_1^2 + x_2 - 1))$) is represented on the left by the deformation of the grid $(\frac{i}{18}, \frac{j}{18})_{0\le i,j\le 18}$. Then we display its approximations with different levels of precision. In this example, the basis functions were chosen as radial Gaussian functions with centers positioned on a regular grid of size N_{basis} and with range $\sigma_{basis} = 3/(5N_{basis})$. The weights were computed directly with the values of the true warping.*

Appendix B. Derivation of the gradient-based sampling criteria.

Proof of Proposition 4.2. Let us first address the case $\eta = 1$ using the notation $\boldsymbol{Z}_{\boldsymbol{x}}^c = \boldsymbol{Z}_{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}}$ with $\boldsymbol{m}_{\boldsymbol{x}} = \nabla m_n(\boldsymbol{x})$. The first step is to expand the criterion as follows:

$$\begin{aligned}\operatorname{var}\left(||\boldsymbol{Z}_{\boldsymbol{x}}||^2\right) &= \operatorname{var}\left(\boldsymbol{Z}_{\boldsymbol{x}}^\top \boldsymbol{Z}_{\boldsymbol{x}}\right) = \operatorname{var}\left(2\boldsymbol{m}_{\boldsymbol{x}}^\top \boldsymbol{Z}_{\boldsymbol{x}}^c + \boldsymbol{Z}_{\boldsymbol{x}}^{c\top} \boldsymbol{Z}_{\boldsymbol{x}}^c\right) \\ &= 4\operatorname{var}\left(\boldsymbol{m}_{\boldsymbol{x}}^\top \boldsymbol{Z}_{\boldsymbol{x}}^c\right) + \underbrace{2\operatorname{cov}\left(\boldsymbol{m}_{\boldsymbol{x}}^\top \boldsymbol{Z}_{\boldsymbol{x}}^c, \boldsymbol{Z}_{\boldsymbol{x}}^{c\top} \boldsymbol{Z}_{\boldsymbol{x}}^c\right)}_{=0 \text{ (nullity of } 3^{\text{th}} \text{ order moments)}} + \operatorname{var}\left(\boldsymbol{Z}_{\boldsymbol{x}}^{c\top} \boldsymbol{Z}_{\boldsymbol{x}}^c\right).\end{aligned}$$

The term $\operatorname{var}(\boldsymbol{Z}_{\boldsymbol{x}}^{c\top} \boldsymbol{Z}_{\boldsymbol{x}}^c)$ can be further expanded as $\boldsymbol{Z}_{\boldsymbol{x}}^c = U_{\boldsymbol{x}} D_{\boldsymbol{x}}^{\frac{1}{2}} \boldsymbol{N}$ with $U_{\boldsymbol{x}}$ an orthogonal matrix, $D_{\boldsymbol{x}}$ the diagonal matrix of eigenvalues, and $\boldsymbol{N}$ a standard Gaussian vector:

$$\operatorname{var}\left(\boldsymbol{Z}^{c\top}_{\boldsymbol{x}} \boldsymbol{Z}^c_{\boldsymbol{x}}\right) = \operatorname{var}\left((U_{\boldsymbol{x}}\boldsymbol{N})^\top D_{\boldsymbol{x}} (U_{\boldsymbol{x}}\boldsymbol{N})\right) = \sum_{i=1}^{d} \lambda_i(\boldsymbol{x})^2 \underbrace{\operatorname{var}\left(N_i^2\right)}_{=2}.$$

For $\eta = 1$, considering the variance of $||\boldsymbol{Z}_{\boldsymbol{x}}||$ in terms of raw moments gives

$$\begin{aligned}\operatorname{var}\left(||\boldsymbol{Z}_{\boldsymbol{x}}||\right) &= \mathbb{E}\left(\boldsymbol{Z}_{\boldsymbol{x}}^\top \boldsymbol{Z}_{\boldsymbol{x}}\right) - \mathbb{E}\left(\sqrt{\boldsymbol{Z}_{\boldsymbol{x}}^\top \boldsymbol{Z}_{\boldsymbol{x}}}\right)^2 \\ &= \boldsymbol{m}_{\boldsymbol{x}}^\top \boldsymbol{m}_{\boldsymbol{x}} + \underbrace{2\boldsymbol{m}_{\boldsymbol{x}}^\top \mathbb{E}\left(\boldsymbol{Z}_{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}}\right)}_{=0} + \sum_{i=1}^{d} \operatorname{var}\left([\boldsymbol{Z}_{\boldsymbol{x}}]_i\right) - \mathbb{E}\left(\sqrt{Q_{\boldsymbol{x}}}\right)^2.\end{aligned}$$

For the proof of $\mathrm{IGNV}_{\eta=1,2}$, we focus on the integrand. We formulate the case $\eta = 2$:

$$\begin{aligned}\mathbb{E}\left(\operatorname{var}\left(Q_{\boldsymbol{u}} \mid \mathcal{A}_n, Y_{\boldsymbol{x}}\right) \middle| \mathcal{A}_n\right) =& 4\mathbb{E}\left(\mathbb{E}\left(\nabla Y_{\boldsymbol{u}} \middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right)^\top \nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \mathbb{E}\left(\nabla Y_{\boldsymbol{u}} \middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right) \middle| \mathcal{A}_n\right) \\ &+ 2\sum \lambda_{i,\boldsymbol{x}}(\boldsymbol{u})^2. \qquad (25)\end{aligned}$$

We get the result with $\mathbb{E}(\nabla Y_{\boldsymbol{u}} | \mathcal{A}_n, Y_{\boldsymbol{x}}) = \nabla \boldsymbol{m}_n(\boldsymbol{u}) + \frac{Y_{\boldsymbol{x}} - m_n(\boldsymbol{x})}{c_n(\boldsymbol{x},\boldsymbol{x})} \kappa_n(\boldsymbol{u},\boldsymbol{x})$.

For $\eta = 1$, we obtain

$$\begin{aligned}\mathbb{E}\left(\operatorname{var}\left(\sqrt{Q_{\boldsymbol{u}}} \middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right) \middle| \mathcal{A}_n\right) =& \mathbb{E}\left(||\mathbb{E}\left(\nabla Y_{\boldsymbol{u}} \middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right)||^2 \middle| \mathcal{A}_n\right) + \operatorname{tr}\left(\nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u})\right) \\ &- \mathbb{E}\left(\mathbb{E}\left(\sqrt{Q_{\boldsymbol{u}}} | \mathcal{A}_n, Y_{\boldsymbol{x}}\right)^2 \middle| \mathcal{A}_n\right). \qquad (26)\end{aligned}$$

Finally, replacing $\mathbb{E}(\nabla Y_{\boldsymbol{u}} | \mathcal{A}_n, Y_{\boldsymbol{x}})$ by its analytic formula gives the result. ∎

REFERENCES

[1] E. B. Anderes and M. L. Stein, *Estimating deformations of isotropic gaussian random fields on the plane*, Ann. Statist., 36 (2008), pp. 719–741.

[2] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez, *Sequential design of computer experiments for the estimation of a probability of failure*, Stat. Comput., 22 (2012), pp. 773–793.

[3] D. Brillinger, *The identification of a particular nonlinear time series system*, Biometrika, 64 (1977), pp. 509–515.
[4] C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Y. Richet, *Fast kriging-based stepwise uncertainty reduction with application to the identification of an excursion set*, Technometrics, 56 (2014), pp. 455–465.
[5] C. Chevalier, D. Ginsbourger, and X. Emery, *Corrected kriging update formulae for batch-sequential data assimilation*, in Mathematics of Planet Earth, Springer, New York, 2014, pp. 119–122.
[6] T. Choi, J. Q. Shi, and B. Wang, *A gaussian process regression approach to a single-index model*, J. Nonparametr. Stat., 23 (2011), pp. 21–36.
[7] Y. Deville, D. Ginsbourger, and O. Roustant, *Package 'kergp'*, 2015.
[8] J. Doob, *Stochastic Processes*, John Wiley & Sons, New York, 1953.
[9] P. Duchesne and P. De Micheaux, *Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods*, Comput. Statist. Data Anal., 54 (2010), pp. 858–862.
[10] D. Dupuy, C. Helbert, and J. Franco, *DiceDesign and DiceEval: Two R packages for design and analysis of computer experiments*, J. Stat. Softw., 65 (2015), pp. 1–38, http://www.jstatsoft.org/v65/i11/.
[11] N. Durrande, D. Ginsbourger, and O. Roustant, *Additive covariance kernels for high-dimensional gaussian process modeling*, in Ann. Fac. Sci. Toulouse, 21 (2012), pp. 481–499.
[12] M. Gibbs, *Bayesian Gaussian Processes for Regression and Classification*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1997.
[13] D. Ginsbourger, O. Roustant, and N. Durrande, *On degeneracy and invariances of random fields paths with applications in gaussian process modelling*, J. Statist. Plann. Inference, 170 (2016), pp. 117–128.
[14] D. Ginsbourger, O. Roustant, D. Schuhmacher, N. Durrande, and N. Lenz, *On ANOVA decompositions of kernels and gaussian random field paths*, in Monte Carlo and Quasi-Monte Carlo Methods, Springer, New York, 2016, pp. 315–330.
[15] R. B. Gramacy, *Bayesian Treed Gaussian Process Models*, Ph.D. thesis, University of California Santa Cruz, 2005.
[16] R. B. Gramacy, *TGP: An R package for bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models*, J. Stat. Softw., 19 (2007), https://ideas.repec.org/a/jss/jstsof/v019i09.html.
[17] R. B. Gramacy and H. K. H. Lee, *Bayesian treed gaussian process models with an application to computer modeling*, J. Amer. Statist. Assoc., 103 (2008), pp. 1119–1130, http://dx.doi.org/10.1198/016214508000000689.
[18] R. B. Gramacy and H. K. H. Lee, *Adaptive design and analysis of supercomputer experiments*, Technometrics, 51 (2009), pp. 130–145.
[19] R. B. Gramacy and H. Lian, *Gaussian process single-index models as emulators for computer experiments*, Technometrics, 54 (2012), pp. 30–41.
[20] R. B. Gramacy and M. A. Taddy, *Categorical inputs, sensitivity analysis, optimization and importance tempering with TGP version 2, an R package for treed gaussian process models*, J. Stat. Softw., 33 (2010), https://ideas.repec.org/a/jss/jstsof/v033i06.html.
[21] D. R. Jones, M. Schonlau, and J. William, *Efficient global optimization of expensive black-box functions*, J. Global Optim., 13 (1998), pp. 455–492.
[22] D. J. C. MacKay, *Introduction to gaussian processes*, NATO ASI Ser. F Comput. Systems Sci., 168 (1998), pp. 133–166.
[23] J. Mockus, *Application of Bayesian approach to numerical methods of global and stochastic optimization*, J. Global Optim., 4 (1994), pp. 347–365.
[24] C. Paciorek, *Nonstationary Gaussian Processes for Regression and Spatial Modelling*, Ph.D. thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, 2003.
[25] C. Paciorek and M. Schervish, *Nonstationary covariance functions for gaussian process regression*, Adv. Neural Inf. Process. Syst., 16 (2004), pp. 273–280.
[26] E. Padonou and O. Roustant, *Polar gaussian processes and experimental designs in circular domains*, SIAM/ASA J. Uncertain. Quantif., 4 (2016), pp. 1014–1033.

[27] J.-S. Park and J. Baek, *Efficient computation of maximum likelihood estimator in a spatial linear model with power exponential covariogram*, Comput. Geosci., 27 (2001), pp. 1–7.
[28] F. Perales, F. Dubois, Y. Monerie, B. Piar, and L. Stainier, *A nonsmooth contact dynamics-based multi-domain solver, code coupling (Xper) and application to fracture*, Eur. J. Comput. Mech., 19 (2010), pp. 389–417.
[29] V. Picheny, D. Ginsbourger, O. Roustant, R. T. Haftka, and N.-H. Kim, *Adaptive designs of experiments for accurate approximation of target regions*, J. Mech. Design, 132 (2010).
[30] L. Pronzato and W. G. Müller, *Design of computer experiments: Space filling and beyond*, Stat. Comput., 22 (2011), pp. 681–701, http://dx.doi.org/10.1007/s11222-011-9242-3.
[31] J. Quiñonero-Candela and C. E. Rasmussen, *A unifying view of sparse approximate gaussian process regression*, J. Mach. Learn. Res., 6 (2005), pp. 1939–1959.
[32] P. Ranjan, D. Bingham, and G. Michailidis, *Sequential experiment design for contour estimation from complex computer codes*, Technometrics, 50 (2008), pp. 527–541.
[33] C. R. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, UK, 2006.
[34] S. E. Rogers, M. J. Aftosmis, S. A. Pandya, N. M. Chaderjian, E. Tejnil, and J. U. Ahmad, *Automated CFD Parameter Studies on Distributed Parallel Computers*, AIAA paper 4229, AIAA, 2003.
[35] J. Rougier, *A Representation Theorem for Stochastic Processes with Separable Covariance Functions, and Its Implications for Emulation*, arXiv:1702.05599 [math.ST], 2017.
[36] O. Roustant, D. Ginsbourger, and Y. Deville, *DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by Kriging-based metamodelling and optimization*, J. Stat. Softw., 51 (2012), pp. 1–55, http://www.jstatsoft.org/v51/i01/.
[37] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statist. Sci., (1989), pp. 409–423.
[38] P. D. Sampson and P. Guttorp, *Nonparametric Estimation of Nonstationary Spatial Covariance Structure*, J. Amer. Statist. Assoc., 87 (1992), pp. 108–119.
[39] T. J. Santner, B. J. Williams, and W. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2003.
[40] M. Scheuerer, *A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis*, Ph.D. thesis, Georg-August-Universität, Göttingen, 2009.
[41] J. Snoek, K. Swersky, R. S. Zemel, and R. P. Adams, *Input warping for bayesian optimization of non-stationary functions*, in Proceedings of ICML, 2014, pp. 1674–1682.
[42] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, *Information-theoretic regret bounds for gaussian process optimization in the bandit setting*, IEEE Trans. Inform. Theory, 58 (2012), pp. 3250–3265.
[43] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York, 1999.
[44] E. Vazquez and J. Bect, *Sequential search based on kriging: Convergence analysis of some algorithms*, in Proceedings of the 58th World Statistics Congress of the International Statistical Institute, Dublin, Ireland, 2011, arXiv:1111.3866.
[45] B. J. Williams, T. J. Santner, and W. I. Notz, *Sequential design of computer experiments to minimize integrated response functions*, Statist. Sinica, 10 (2000), pp. 1133–1152, http://www3.stat.sinica.edu.tw/statistica/oldpdf/A10n46.pdf.
[46] Y. Xia, *A multiple-index model and dimension reduction*, J. Amer. Statist. Assoc., 103 (2008), pp. 1631–1640.
[47] Y. Xiong, W. Chen, D. Apley, and X. Ding, *A non-stationary covariance-based kriging method for metamodelling in engineering design*, Internat. J. Numer. Methods Engrg., 71 (2007), pp. 733–756, http://dx.doi.org/10.1002/nme.1969.