Warped Gaussian processes and derivative-based sequential design for functions with heterogeneous variations

Sébastien Marmin, David Ginsbourger, Jean Baccou, Jacques Liandrat

Warped Gaussian processes and derivative-based sequential design for functions with heterogeneous variations

Sébastien Marmin[1,3,4,6,*], David Ginsbourger[2,1,6], Jean Baccou[3,5], Jacques Liandrat[4]

Gaussian process (GP) models have become popular for approximating and exploring non-linear systems based on scarce input/output training samples and on prior hypotheses implicitly done through prior mean and covariance functions. While it is common to make stationarity assumptions and use variance-based criteria for space exploration, in realistic test cases it is not rare that systems under study exhibit an heterogeneous behaviour depending on considered regions of the parameter space. With a class of problems in mind where high variations occur along unknown non-canonical directions, we tackle the problem of uncovering and accommodating non-stationarity in function approximation from two angles: first via a novel class of covariances (called WaMI-GP) that simultaneously generalizes kernels of Multiple Index and of tensorized warped GP models and second, by introducing derivative-based sampling criteria dedicated to the exploration of high variation regions. The novel GP class is investigated both through mathematical analysis and numerical experiments, and it is shown that proposed kernels allow encoding much expressiveness while remaining with a moderate number of parameters to be inferred. On the other hand, and independently of non-stationarity assumptions, we conduct (semi-)analytically derivations for our new variance-based infill sampling criteria relying on a change of focus from the GP to the norm of its associated gradient field. Criteria and GP models are first compared on a mechanical test case taken from nuclear safety studies conducted by IRSN. It is found on this application that some of the proposed sampling criteria including derivatives outperform usual variance-based criteria in the case of a stationary GP model, but that it is even better to use standard variance-based criteria with the proposed novel class of covariances. Comparisons are also done with the Treed Gaussian Processes (TGP) both on this application and on a three-dimensional 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 initial designs it is outperformed by WaMI-GP in the sequential set up.

Keywords: Non-stationary kernels, Infill sampling criteria, Computer Experiments.

*Corresponding author: sebastien.marmin@irsn.fr

1. University of Bern, IMSV, Alpeneggstrasse 22, CH-3012 Bern, Switzerland.
2. Idiap Research Institute, UQOD group, Centre du Parc, rue Marconi 19, CH-1920 Martigny, Switzerland
3. Institut de Radioprotection et de Sûreté Nucléaire, PSN-RES/SEMIA/LIMAR, CEA Cadarache, 13114 Saint-Paul-lez-Durance, France.
4. Centrale Marseille, I2M, UMR 7373, CNRS, Université Aix-Marseille, 13453 Marseille, France.
5. Laboratoire de Micromécanique et d'Intégrité des Structures, IRSN-CNRS-UMII, B.P. 3, 13115 Saint-Paul-lez-Durance, France.
6. MASCOT-NUM Research Group - Méthodes d'Analyse Stochastique des Codes et Traitements Numériques.

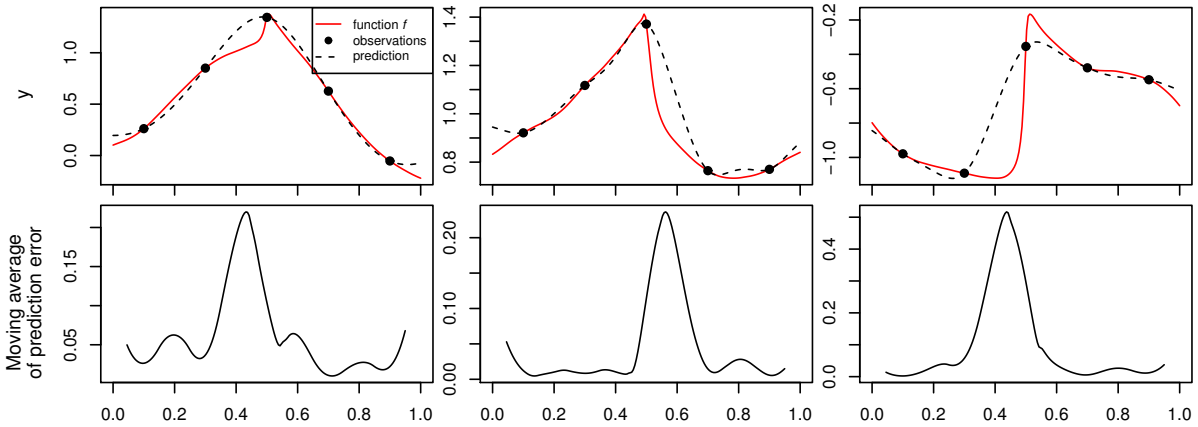

Figure 1: Concentration of the prediction error around high variation zones (empirical assessment). Top: functions with high variation zones (obtained by generating sample paths of a simple non-stationary GP) with prediction curves from five evaluations. A classical interpolating model is used (GP model with stationary covariance of type Matérn $\nu = \frac{5}{2}$). Bottom: absolute differences between predictions and true values, averaged on a moving window of width $\frac{1}{10}$.

1 Introduction

Many experimental systems abruptly change regime: in material science, with percolation threshold of porous media, in epidemiology, with outbreak of a pathogen according to uncertain characteristics of a population, in thermodynamics with phase transition, etc. This situation is also often encountered in nuclear safety analysis in which, to take one example, slight variations in input parameters of computer codes may strongly impact responses quantifying system safety due to a steep transition between competing mechanical phenomena. Let us assume that we aim to study 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 for instance by changing magnitude of the gradient norm depending on regions of the input space or, to take one alternative viewpoint, by spatially-varying main local frequencies. Here we will informally refer to such f's as "functions with heterogeneous variations". As illustrated in figure 1, regions with abrupt changes in function values can lead to increased prediction errors. Allocating more evaluation in these regions is a natural idea. However, practitioners are often in the situation that they only know the existence of such heterogeneities without much precise information regarding their location, shape, or orientation. In the context of expensive evaluations it thus makes sense to appeal to modelling and sampling approaches that allow better approximating f in those regions based on data. 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 most promising next evaluation point(s). Evaluations of f at points deemed most promising and surrogate model updates are then repeated until stopping condition is met, e.g. depletion of the evaluation budget.

Gaussian Process (GP) models are popular surrogate models, and have become a standard in the design and analysis of computer experiments (see e.g. [38], [21] and [40]). They consist in assuming that the unknown objective function f is a sample path of a *prior* GP Y indexed by the source space of f. The term *prior* refers to probability distributions assumed before any evaluation, to oppose to *posterior* distributions that take some evaluation results into account. The free determination of a prior makes GP models versatile for guiding additional evaluations

according to a user defined goal and for integrating practitioner's initial knowledge on f. The quality of the model depends both on the data set and on the adequacy between the prior GP Y and the actual f to be predicted. Adapting the prior covariance of Y to specific classes of objective functions f has inspired a lot of research results. For example, for objective functions with a better representation in polar coordinate, [27] 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 [14], and for functions with a sparse ANOVA decomposition [11] [15]. In the absence of such specific prior assumption of f, it is common to take stationary covariance functions [44]. A GP model with a stationary prior covariance means that in absence of evaluation, the distribution of outputs $(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'})^\top$, for every pair $(\boldsymbol{x}, \boldsymbol{x}')$ in the input space, depends only on the difference $\boldsymbol{x} - \boldsymbol{x}'$. Among stationary kernels, the Matérn class is quite popular as it allows tuning the order of (almost sure) differentiability of associated GP realizations. Note that both tensor product Matérn kernels (e.g. in [37]) and their radial counterparts (such as in [34]) 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 [36] with a focus on the tensor product (or "separability") property.

Yet, when f is known to possess heterogeneous variations, it is sensible to depart from the stationary hypothesis from the start and consider non-stationary prior covariances that account for this property. Among various proposals from non-stationary GP modelling, we can cite convolution methods, see [26] [13], or input space warping approaches ([39]). In this last approach, the non-stationary GP comes from the chaining of a GP with a warping function of D, corresponding to what is often referred to as a *change of time* in stochastic process theory. A strongly consistent approach for estimating warpings of a bivariate isotropic GP from dense evaluations of a single (warped) realisation is provided by [1]. In contrast, challenges considered here rather call for warping estimation from scarce evaluations in order to build appropriate non-stationary surrogates for f's with arbitrary d-dimensional source space. Gibbs [13] tackled this problem using parametric warpings relying on linear combinations of basis functions. Following this idea, Xiong et al. [48] drastically reduced the number of parameters by taking tensor products of univariate warpings. An additional well recognised model around the sphere of GP modelling, Treed Gaussian Process (TGP) [16], is a spatial assembly of different GPs. Based on partitioning input space, its current implementation divides D into parallelepipeds on which individual GPs are defined. While this method is very flexible and allows by construction accounting for heterogeneous variations, it steps out the GP paradigm in the strict sense and hence does not benefit from some of its convenient properties. In the present context of small data sets and heterogeneous variations driven by unknown directions, GP surrogate models are needed that enjoy the sparsity of Xiong et al.'s axial warping 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* such as GP-SIM [19] and more generally to so-called *multiple index models* [47]. Our proposed kernel class is underlain by these considerations, as detailed in next sections.

From a different perspective, GP models 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 [45] and hence to eventually learn high-variation regions, in stationary cases it is done in a non-adaptive way as then the prediction variance 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 [46, 33, 31, 2, 4]. On a different note, adaptive design criteria have been used for global optimization (See notably [24, 21, 43]). Notably, input warping has been recently shown to improve Bayesian optimization in non-stationary cases [42].

Here we tackle the problem of learning functions with heterogeneous variations along unknown directions from a limited number of evaluations both from the modelling and the sequential design points of view. Contributions are organised in two parts. First we introduce the Warped Multiple Index (WaMI) GP model. It relies on a new family of non-stationary covariance kernels that combines features from Multiple Index GPs and tensorized warpings. A nice aspect of this kernel family is the number of hyperparameters that increase affinely with the dimension (with slope 1). Besides this, the model can incorporate any orientation of heterogeneous variations. Regarding sampling aspects, we develop targeted criteria for adaptively sampling functions with heterogeneous variations. Based on GP gradient norms, they make a trade-off between space-fillingness and intensifying exploration in high variation regions of f.

We apply these contributions on functions arising 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, our proposed 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 the criteria (MSE and IMSE versus introduced derivative based criteria) and GP models (stationary versus WaMI). Best results are obtained using WaMI-GP (and to a lesser extent, TGP) combined with the classical variance-based criteria, followed by a combination of a stationary GP model combined with one of the proposed gradient-based criteria. In the second test case, a three-dimensional fluid dynamics application from NASA that was used in seminal article about TGP, TGP remains the best model at fixed space-filling design of experiments, but when both methods are combined with sequential design (using the MSE criterion here) WaMI-GP outperforms TGP by successfully detecting high variation regions and attributing them higher prediction variance values.

The paper is organised as follows: Section 2 is devoted to an overview on GP models with a focus on non-stationary models, and their use for sequential design of experiments. Then we introduce and investigate the WaMI-GP model in Section 3, followed by several proposals of derivative-based criteria for the exploration of high-variation regions in Section 4. Finally, experimental results and comparisons with classical approaches based on the two engineering test cases are presented in Section 5.

2 State-of-the-art

2.1 Stochastic modelling for the emulation of computer experiments

2.1.1 Overview of Gaussian process modelling basics

In GP modelling, 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* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ (here a finite-dimensional vector). In this article, we adopt an empirical Bayes viewpoint, meaning in the present context 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 based on data. In other words, given $\boldsymbol{\theta}$, for all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in D$, the corresponding response vector $(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_p))^\top$ (standing for values of the objective functions at those points) is *a priori* distributed as a multivariate normal distribution $\mathcal{N}((m_{\boldsymbol{\theta}_1}(\boldsymbol{x}_i))_{i=1,\ldots,p}^\top, (c_{\boldsymbol{\theta}_2}(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1,\ldots,p})$. Estimators for $\boldsymbol{\theta}$ can notably be defined by cross-validation minimization or by maximum likelihood (see [28]) and then simply plugged-in. While in such case it is customary to distinguish the es-

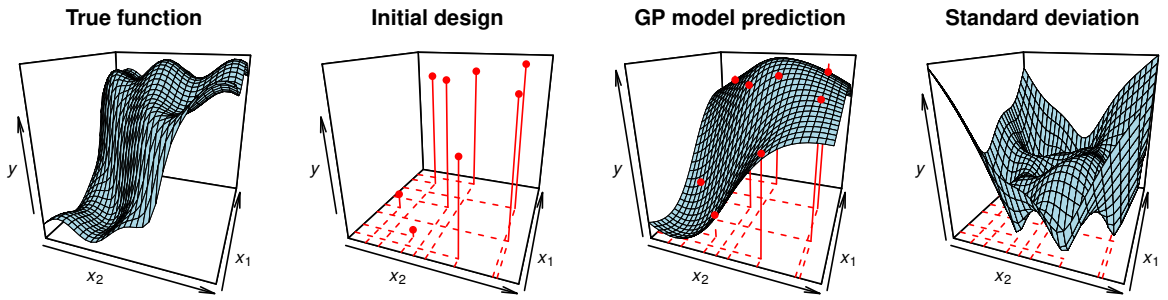

Figure 2: Stationary GP modelling of a toy function (equation (3)). **From left to right:** representation of f; locations and responses of eight initial evaluations; posterior mean of the Gaussian process; posterior standard deviation.

timate from a fixed value by putting hat on $\boldsymbol{\theta}$, here estimated parameters are often implicit for the sake of readability. Also, numerical details about mean and covariance parameter estimation can be found in [37] and are omitted here for conciseness.

Now given n arbitrary points $\boldsymbol{x}_1, \dots \boldsymbol{x}_n \in D$ and observed values $(y(\boldsymbol{x}_1), \dots, y(\boldsymbol{x}_p))$ of Y at those points, the so-called *kriging formulae* and the underlying posterior GP model are obtained by conditioning Y on the event $\mathcal{A}_n = \{Y(\boldsymbol{x}_1) = y_1, \dots, Y(\boldsymbol{x}_n) = y_n\}$: $c_n : (\boldsymbol{x}, \boldsymbol{x}') \to \mathrm{cov}\left(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'} | \mathcal{A}_n\right)$ informing on the prediction uncertainty, including the posterior standard deviation defined as $\boldsymbol{x} \to \sqrt{c_n(\boldsymbol{x}, \boldsymbol{x})}$. We have:

$$m_n(\boldsymbol{x}) = \mathbb{E}\left(Y_{\boldsymbol{x}} | \mathcal{A}_n\right) = m(\boldsymbol{x}) + \boldsymbol{c}_n(\boldsymbol{x})^\top C_n^{-1}(\boldsymbol{y}_{1:n} - (m(\boldsymbol{x}_i))_{i=1,\dots,n}^\top), \text{ and} \tag{1}$$

$$c_n(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{cov}\left(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'} | \mathcal{A}_n\right) = c(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{c}_n(\boldsymbol{x})^\top C_n^{-1} \boldsymbol{c}_n(\boldsymbol{x}'), \tag{2}$$

where $\boldsymbol{y}_{1:n} = (y_i)_{i=1,\dots,n}^\top$, $\boldsymbol{c}_n(\boldsymbol{x}) = (c(\boldsymbol{x}, \boldsymbol{x}_i))_{1 \le i \le n}^\top$, and $C_n = (c(\boldsymbol{x}_i, \boldsymbol{x}_j))_{1 \le i,j \le n}$ (assumed non-singular here). A bivariate stationary GP model is illustrated in Figure 2, 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 8 observations determined by a space filling algorithm : random LHS design optimized with a maximin 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 [37]). The covariance kernel used in Figure 2 is a stationary Matérn (See [44, 34]) with $\nu = 5/2$ and $\boldsymbol{\theta} \in (\mathbb{R}_+ \backslash \{0\})^{d+1}$:

$$c_{\boldsymbol{\theta}}\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \theta_{d+1}\left(1 + \sqrt{5}h + \frac{5}{3}h^2\right)\exp\left(-\sqrt{5}h\right), \text{ where } h = \sqrt{\sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\theta_i^2}}. \tag{4}$$

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 non-stationary model may improve the model by fitting the heterogeneous behaviour of f as discussed below.

2.1.2 Non-stationary approaches: from warped GP to TGP

Non-stationary GP models are already widely applied. It often consists in injecting prior knowledge about spatial-dependency. There are several approaches to achieve that. A trivial way is

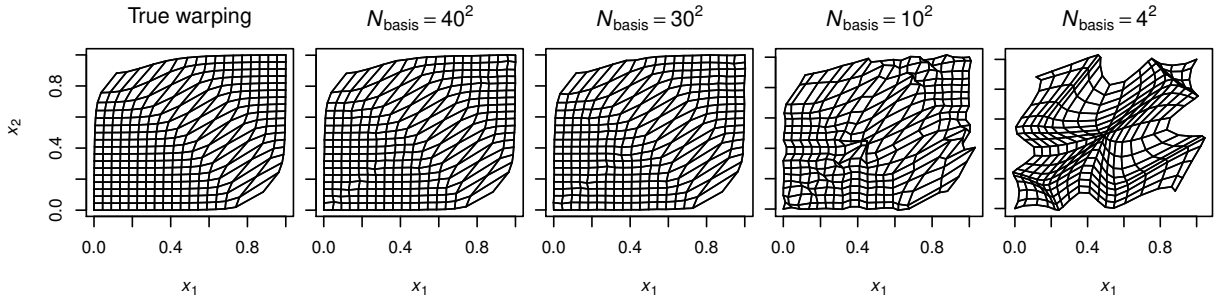

Figure 3: Warping approximation with Gibbs' method, for different number of basis functions. An arbitrary warping (equation (5)), is represented on the left by the deformation of the grid $(\frac{i}{18}, \frac{j}{18})_{i,j=0,\dots,18}$. Then we display its approximations with different levels of precision.

based on vertical scaling (see e.g. [23] for a formulation): as $c(\boldsymbol{x}, \boldsymbol{x}) = \sigma^2$ is constant for any stationary covariance c, replacing $c(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 R(\boldsymbol{x}, \boldsymbol{x}')$ (where R is the corresponding correlation function) by $c(\boldsymbol{x}, \boldsymbol{x}') = \sigma(\boldsymbol{x})\sigma(\boldsymbol{x}')R(\boldsymbol{x}, \boldsymbol{x}')$ where σ is a non-negative function that creates a non-stationary covariance. Moreover, Paciorek and Schervish [26] address the issue of non-stationarity by extending a convolution method proposed in [13] from a squared exponential kernel to any covariance structure. Warping stationary GPs for creating non-stationary GPs is also a common method (See, e.g., [39]). In this approach, called the non-linear map method, the non-stationary covariance function c is derived from the GP $\boldsymbol{x} \to Y_{\boldsymbol{x}} = Z_{T(\boldsymbol{x})}$, with Z a generally 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 obtained by chaining the stationary covariance with $T \otimes T$; i.e. $\forall \boldsymbol{x}, \boldsymbol{x}' \in D$, $c(\boldsymbol{x}, \boldsymbol{x}') = k(T(\boldsymbol{x}), T(\boldsymbol{x}'))$.

The estimation of T from data is a difficult problem. The non-linear map method flexibility is also challenging for the estimation of T among the set of injections on D. A first restriction, implicitly assumed in almost all applications, is to consider only continuous bijections. The estimation of T is also often simplified to a finite dimension problem taking $T = T_{\boldsymbol{\tau}}$, with $\boldsymbol{\tau}$ a parameter vector. Gibbs' method [13], for example, formulates $T_{\boldsymbol{\tau}}$ as a multidimensional integral of non-negative density functions, that ensure its bijectivity and continuity:

$$T_\tau(\boldsymbol{x}) = \boldsymbol{x}_0 + \left(\int_{P_{01}} g_i(\boldsymbol{u}) \mathrm{d}s \right)^\top_{i=1,\dots,d},$$

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. The estimation of T reduces to the estimation of a finite number of weights. We see in Figure 3 how this method allows an approximation of a given deformation

$$T_0(\boldsymbol{x}) = \boldsymbol{x} + 1/10 \arctan(30(x_1^2 + x_2 - 1)). \tag{5}$$

In this example, the basis functions were chosen as uncorrelated Gaussian functions with centres positioned on a regular grid of size N_{basis} and with range $\sigma_{\text{basis}} = 3/(5N_{\text{basis}})$. The weights were computed directly with the values of the deformation at the centres of the basis functions.

We observe a degradation of the warping approximation with decreasing numbers of parameters to estimate: with a grid of 16 basis functions, i.e. requiring a computation of 32 weights in dimension 2, the approximation fails despite a relatively large number of parameters. Here about 100 basis functions are needed to capture the non-stationarity in the whole domain. This reduces the applicability of the method in contexts with drastically numbers of evaluations. Note that keeping the same level of spatial precision, say r basis functions for each direction, the number

dr^d of weights increases rapidly with d. Therefore an effort has been done to reduce the number of parameters while preserving some flexibility. e.g. with the axial warping method [48].

In this method, it is assumed that for $\boldsymbol{x} \in D$, $T(\boldsymbol{x}) = (T_i(x_i))_{i=1,\dots,d}^\top$, with $(T_i)_{i=1,\dots,d}$ continuous univariate bijections. They take for T_i, $i = 1, \dots, p$, piecewise second degree polynomials, with differentiability constraints and nodes placed along the i^{th} dimension. In Figure 4 we display the results of applying this method to the toy function already used in Figure 2, along with the estimated axial warping densities.

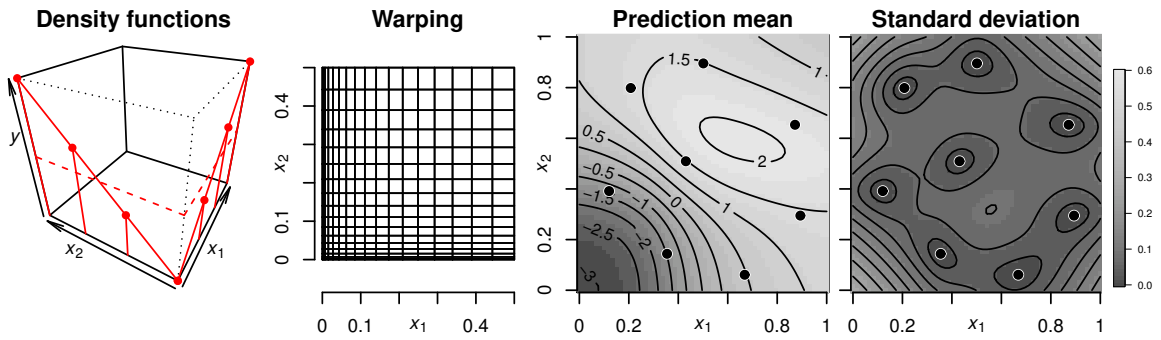

Figure 4: Gaussian process model with axial warping, applied to the running example function. **From left to right:** estimated warping density functions for the axes; corresponding surface warping of $[0,1]^2$ (represented by deformation of a 10×10 regular orthogonal grid); prediction mean and standard deviation.

In some situations, warping only along canonical axis 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 non-canonical 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.

Another strategy for functions with high variation zones, and close to GP modelling, is the (Bayesian) Treed Gaussian Process (TGP). This method is based on partitioning the input space. Different GP models are then associated independently in each partition, allowing highly heterogeneous behaviour across the input space. A strength of this method is that partitions and their number are automatically determined according to the data. The discontinuity resulting from partitioning could sound like a drawback, although it is shown that, due to Bayesian averaging which is not detailed here, it does not increase notably the prediction error. Figure 5 shows the application on the toy function obtained with the R package 'tgp'.

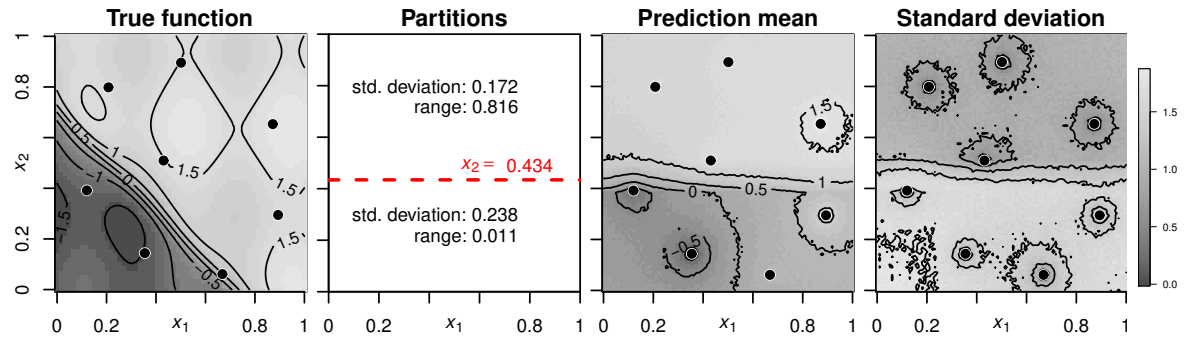

Figure 5: Bayesian treed Gaussian process model. **From left to right:** the objective function with an initial design; a sketch of the input space partition, the red line divides the space in two regions with different range and standard deviation for different GP models, the prediction mean and standard deviation, and the prediction error.

We observe that the TGP model is able to estimate a partition of the input space in two zones. The algorithm aims at discriminating regions of different variation behaviour. Indeed the region $x_2 > 0.465$ appears to have less variations than the region $x_2 \leq 0.465$. The partitions are implemented to be defined in terms of the canonical axes. But one can expect that when adding new evaluations, allowing partitions not parallel to these axes could improve the modelling. 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.

2.1.3 Dimension reduction with the Multiple Index Model

With expensive evaluations, and therefore often small sample size, it is important to keep the number of model parameters moderate. But in most anisotropic GP models, the dimension of $\boldsymbol{\theta}$ increases rapidly with d. Geometric anisotropies in dimension d require to parametrize a rotation ($d(d-1)/$ parameters) and length-scale parameters for each relevant directions (see e.g [34] p. 106 for a presentation of the squared exponential anisotropic GP). To avoid this, one can consider parsimonious multidimensional non-linear regression like the Single Index Model [3]. Gaussian Process-SIM (GP-SIM [6]) is a particular SIM that can be formulated as a stationary, canonical Gaussian process model [19]. In our empirical Bayes context, the model is defined by its prior covariance: a univariate covariance $k_{\boldsymbol{\beta}}$, parametrised by a vector $\boldsymbol{\beta}$, chained with a scalar product with a vector $\boldsymbol{a} \in \mathbb{R}^d$:

$$c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\boldsymbol{\beta}}(\boldsymbol{a}^\top \boldsymbol{x}, \boldsymbol{a}^\top \boldsymbol{x}'). \tag{6}$$

With this covariance function, the dimension of $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{a}\}$ increases affinely in d with slope 1. In empirical Bayesian setting, this model produces constant predictions in all hyperplanes orthogonal to $\boldsymbol{a}$. Relaxing this constraint, the multiple index model is an extension proposed by [47]. It uses a more complex q-variate covariance function for $k_{\boldsymbol{\beta}}$, and extends the scalar product to a matrix product:

$$c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\boldsymbol{\beta}}\left(A\boldsymbol{x}, A\boldsymbol{x}'\right) \tag{7}$$

with $q \in \mathbb{N}\backslash\{0\}$, A a $q \times d$ matrix, and $k_{\boldsymbol{\beta}}$ a parametrized q dimensional definite positive kernel. The main asset of MIM used in this work is its neutrality towards the canonical axis. Indeed, if we create a new data set applying a linear invertible transformation of the input data $\boldsymbol{x} \to Q\boldsymbol{x}$, the posterior distribution with parameters (β, A^*) for the original model (and thus the likelihood and the cross validation) is the same as the posterior distribution of the transformed model with

parameters $(\beta, A^* Q^{-1})$. Any invertible linear pre-treatment of the data leads intrinsically to the same estimation problem, which is not the case for many models that use the canonical axis as an information for prediction.

2.2 Sampling

2.2.1 Principle and classic criteria

Once a GP model has been built from an initial, say space-filling design of N_{ini} evaluations (like —possibly optimized— LHS designs, minimax-distance designs, etc. See [32] for an overview), the sequential design itself is a loop incrementing n, the current number of evaluations $n = N_{\text{ini}} + 1, \ldots, N$. Sequential sampling is typically driven by the optimization of a family of infill criteria J_n coupled with updating model parameters at each iteration. More precisely, in fully sequential settings, the next evaluation $\boldsymbol{x}_{n+1}$ is selected as a point maximizing a criterion J_n:

$$\boldsymbol{x}_{n+1} \in \underset{\boldsymbol{x} \in D}{\operatorname{argmax}}\ J_n(\boldsymbol{x}). \tag{8}$$

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

A first idea for a criterion is to evaluate the point from which the prediction has the highest posterior variance, in order to reduce the uncertainty of the predicting GP. Classical criteria, Mean Squared Error (MSE) and Integrated MSE (IMSE) are based on this idea. These criteria focus on the zones where the uncertainty on the prediction is high, i.e. where the function is still unexplored according to the model. In the case of deterministic evaluations, the MSE criterion can be reformulated in maximin terms by changing the metric on D to the canonical metric of the GP model covariance c_n, i.e. considering the distance $d(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{c_n(\boldsymbol{x}, \boldsymbol{x}) + c_n(\boldsymbol{x}', \boldsymbol{x}') - 2c_n(\boldsymbol{x}, \boldsymbol{x}')}$. Indeed, the MSE criterion is defined as:

$$J_n^{\text{MSE}}(\boldsymbol{x}) = c_n\left(\boldsymbol{x}, \boldsymbol{x}\right). \tag{9}$$

The latter property highlights that MSE maximization amounts to maximizing a minimal distance to available design points, but taking a distance that accounts for covariances given by the model rather than the Euclidean distance. For IMSE, the aim is to reduce the integral of the MSE over the whole domain D. Thus minimizing the IMSE corresponds to look for a point $\boldsymbol{x}$ minimizing the integral of the future MSE if $\boldsymbol{x}$ is added:

$$J_n^{\text{IMSE}}(\boldsymbol{x}) = \int_{\boldsymbol{u} \in D} c_{n,\boldsymbol{x}}\left(\boldsymbol{u}, \boldsymbol{u}\right) \mathbf{d}\boldsymbol{u}, \tag{10}$$

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

Figure 6 illustrates the first step of a sequential sampling procedure on the toy function that exhibits a high variation zone in the vicinity of the line $x_1 + x_2 = 3/4$. The construction procedure is based on the MSE criterion and exploits the three previously recalled modelling (stationary, non-stationary with axial warping and TGP). As proposals for next evaluations, three MSE maxima can be seen that depend on the model for the running example function. Note that TGP model is not strictly speaking a GP model, but the notion of MSE criterion can still be extended to it.

Non-stationary modelling therefore appears to be a promising choice to tackle the problem of designing experiments in the case of objective functions exhibiting heterogeneous variations.

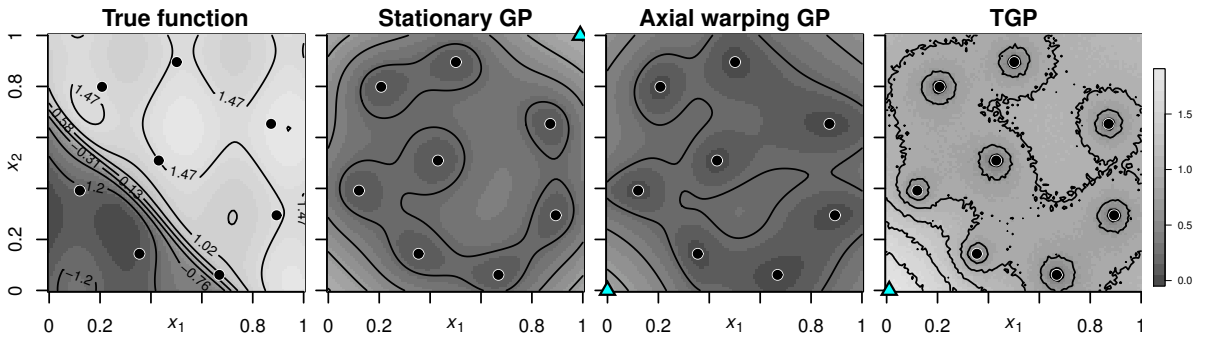

Figure 6: Different GP models of a function $f : \boldsymbol{x} \in [0,1]^2 \to \frac{\sin(15x_1)+\cos(10x_2)}{5} + \arctan\left(\frac{20(x_1+x_2)-15}{2}\right)$. The different models are stationary anisotropic, Xiong's axial approach, and treed GP. For each method, we see the first and last step of the sequential design of experiments, displaying the MSE criterion, the selected point for the next evaluation (blue triangle), and the absolute difference with the real function.

However, as mentioned in Section 2.1.2, existing methods appear to not be fully satisfactory for reducing evaluation budget in some situations. We introduce in the next section a new class of non-stationary models that allows different directions of deformation (other than canonical) while limiting the number of parameters.

3 WaMI-GP: a multiple index model with tensorial deformations

We now present WaMI-GP, a novel of GP models that is dedicated to multivariate functions with heterogeneous variations along non-canonical axes. While this class involves a number of parameters in order to describe the axes as well as the deformations coming into play, their cardinality is kept moderate thanks to the tensorial nature of the involved warping functions. In this section we first introduce the model and gradually illustrate its flexibility, then we provide and prove some of its important properties and 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 here on the covariance kernel of the proposed GP class as, without loss of generality, the GP mean is assumed constant here and in the following.

Definition 1 (WaMI kernel: combining tensorial deformations and multiple index modelling)**.** *Let $q \in \mathbb{N}\backslash\{0\}$, $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q]^\top \in \mathbb{R}^{q\times d}$, $T_i(\cdot, \tau_i) : \mathbb{R} \mapsto \mathbb{R}$ be functions parametrized by $\boldsymbol{\tau}_i$ $(i = 1, \ldots, q)$ 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 vector of parameters containing A, the $\boldsymbol{\tau}_i$'s and $\boldsymbol{\beta}$, we define the associated WaMI (Warped Multiple Index) kernel on D by*

$$c_{\boldsymbol{\theta}} : (\boldsymbol{x}, \boldsymbol{x}') \in D \times D \to c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\boldsymbol{\beta}}\left(\left(T_i(\boldsymbol{a}_i^\top \boldsymbol{x}; \boldsymbol{\tau}_i)\right)_{i=1,\ldots,q}, \left(T_i(\boldsymbol{a}_i^\top \boldsymbol{x}'; \boldsymbol{\tau}_i)\right)_{i=1,\ldots,q}\right). \quad (11)$$

While the symmetric definite positiveness (in the wide sense) of the WaMI kernel is 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 Section 3.3 after some examples. Before we get there, let us already get a feeling of what the WaMI kernel means seen from both individual points of view of axial warping and Multiple Index Modelling (MIM):

1. from the point of view of the axial warping method, we allow non-canonical 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 accomodates dimension reduction (and thus reducing the number of axial warpings) in case $q < d$.

2. from the MIM perspective, we introduce non-stationarity into the covariance by applying non-linear deformations to the result of each scalar product $\boldsymbol{a}_i^\top \boldsymbol{x}$, $(\boldsymbol{a}_i)_{i=1,\dots,q}$.

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 be easily generalized to cases where the warpings are not scalar but rather defined on subspaces of $\mathbb{R}^q$. With our parametrization the total number of parameters is $qd + \#\boldsymbol{\beta} + \sum_{i=1}^{q} \#\boldsymbol{\tau}_i$ where $\#\boldsymbol{\alpha}$ stands for the cardinality of $\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is an arbitrary parameter.

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 type (See equation (4), i.e. with $\nu = 5/2$).

Stationary subcase. Let us first illustrate the case where all univariate deformations are the identity. In Figure 7 we illustrate the warped space (here the overall warping amounts to A), the WaMI kernel and corresponding GP sample paths with

$$A = \begin{pmatrix} 5 & 10 \\ 7.5 & 2.5 \end{pmatrix}. \tag{12}$$

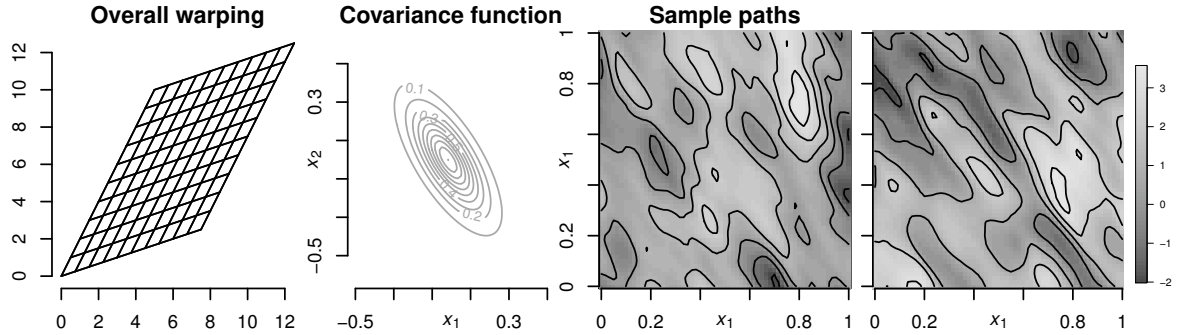

Figure 7: WaMI-GP model in the case of a stationary base kernel and no axial deformation. From left to right: warping represented by mapping of the grid $(\frac{i}{10}, \frac{j}{10})_{i,j=0,\dots,10}$, the covariance function $c(\cdot, (0,0)^\top)$, and three corresponding WaMI-GP realizations.

Note that in case of an isotropic base kernel k_β, the first eigen vectors of $AA^\top$, ordered increasingly by their eigen values, give the directions of high variations appearing in the sample paths. This simple property can be used in a step-by-step parameter estimation procedure for choosing directions in which it is a priority to unlock non-stationarity.

Axial warping subcase. Before combining the effect of a linear transformation and a tensorial warping, we now illustrate the case of axial deformations alone (see Figure 8). 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 cumulative distribution function (CDF) of a beta distribution. It provides a relatively diverse family of non-linear deformations of a segment with only two shape parameters δ_1, δ_2. This function has been also used in other situations for defining univariate deformations, e.g. in [42].

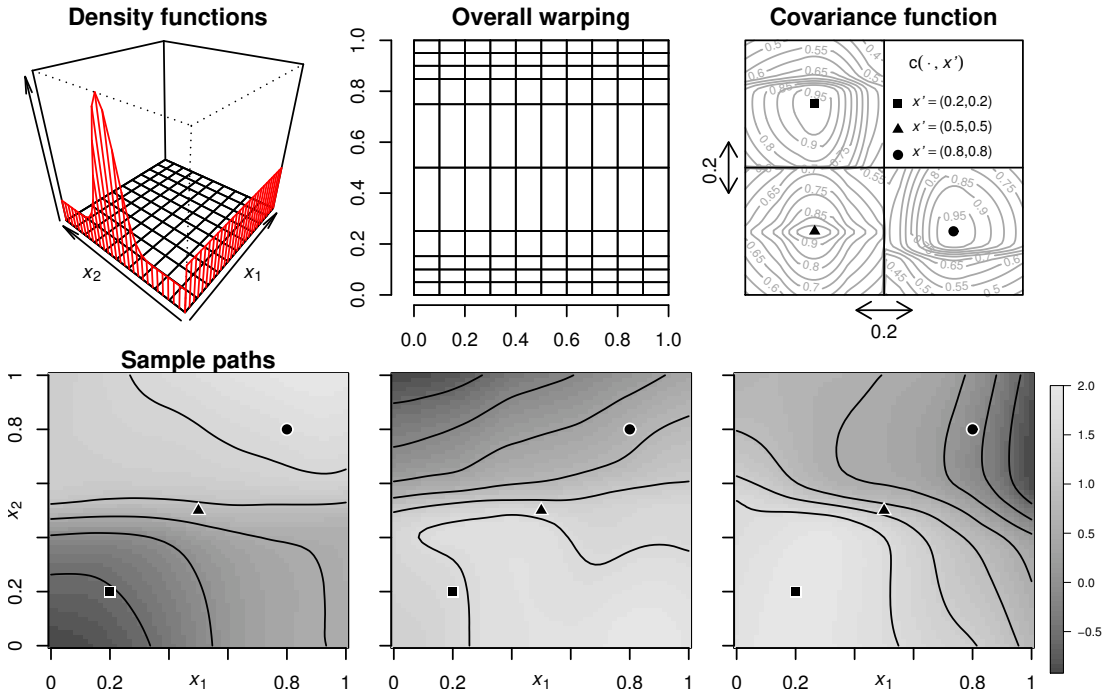

Figure 8: WaMI-GP with axial deformations. From left to right: density function of the deformations in each direction, warping of the grid $(\frac{i}{10}, \frac{j}{10})_{i,j=0,\dots,10}$, the covariance function $c(\cdot, \boldsymbol{x}')$) for different values of $\boldsymbol{x}'$, and two corresponding WaMI GP realisations.

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 non-canonical orientation and two high variation zones. Having a neutral parametrisation towards canonical axes is the key idea for estimating arbitrary directions of heterogeneous variations. We now take

$$A = \begin{pmatrix} \cos(\pi/12) & -\sin(\pi/12) \\ \sin(\pi/12) & \cos(\pi/12) \end{pmatrix}. \tag{13}$$

In addition, we take here T_2 with two ridges,

$$T_2 = x + I(2x; 15, 15) + I(2(x - 1/2); 15, 15), \tag{14}$$

one creates two variation zones. One can observe in Figure 9 the links between high variation regions of realisations and the corresponding overall warping appearing in the corresponding WaMI kernel.

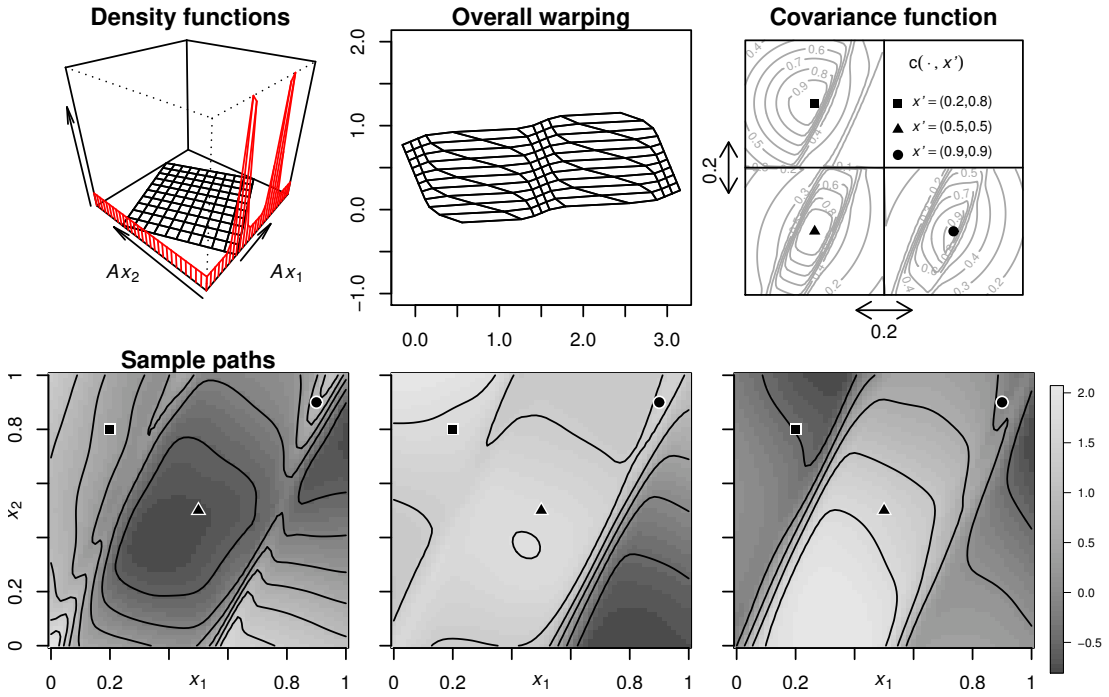

Figure 9: Example of a WaMI-GP with two regions of high variations. From left to right: density functions of the deformations in each direction after the linear transformation, warping of the grid $(\frac{i}{10}, \frac{j}{10})_{i,j=0,\dots,10}$, the covariance function $c(\cdot, \boldsymbol{x}')$ for different values of $\boldsymbol{x}'$, and two corresponding WaMI-GP realizations.

3.3 Strict positive-definiteness and differentiability properties

We prove here properties about the WaMI kernel and the associated (centred) 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 1. ***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 equation (11) is strictly positive definite.*

Proof. Assuming the existence of $\boldsymbol{z}, \boldsymbol{z}' \in D$, with $T(\boldsymbol{z}) = T(\boldsymbol{z}')$, gives $\forall i = 1, \dots, p,\ T_i(z_i; \boldsymbol{\tau}_i) = T_i(z_i'; \boldsymbol{\tau}_i)$ and thus $\boldsymbol{z} = \boldsymbol{z}'$. Moreover, as the rank of A equals the number of its columns, A is injective. So the composition of $\bigotimes_{i=1}^p T_i$ ($\bigotimes$ refeers to tensor product) with $\boldsymbol{x} \to A\boldsymbol{x}$ is injective. Finally the positive definiteness of k_β is conserved by injective chaining. □

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

$$\lim_{h \to 0} \left[\mathbb{E} \left(\left(\frac{Z_{\boldsymbol{x}+h\boldsymbol{e}_i} - Z_{\boldsymbol{x}}}{h} - Z_i^{(1)} \right)^2 \right) \right] = 0. \tag{15}$$

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

Proposition 2. ***Mean-squared differentiability.*** *The centred Gaussian process with the covariance c defined in (11) is mean-squared differentiable (i.e. has mean-squared derivaritives in all canonical directions) under the following conditions:*

- *For all $i \in \{1, \dots, q\}$, $T_i(\cdot; \boldsymbol{\tau}_i)$ have regularity C^1 on $\mathbb{R}$. We denote by $T_i'(\cdot; \boldsymbol{\tau}_i)$ their derivatives.*
- *For all $j, j' \in \{1, \dots, q\}$ and $\boldsymbol{u} \in \mathbb{R}^q$, $\left.\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{(\boldsymbol{u},\boldsymbol{u})}$ exists and is finite.*

Proof. The tensor product T of the $T_i(\cdot; \boldsymbol{\tau}_i)$ functions is also C^1 on $\mathbb{R}^d$. Using the regularity of k_β and T, the chain rule applied to Equation (11) gives that $\forall \boldsymbol{x} \in D$, $\left.\frac{\partial c(\boldsymbol{u},\boldsymbol{u}')}{\partial u_i \partial u'_i}\right|_{(\boldsymbol{x},\boldsymbol{x})}$ exists and is finite. Thus the corresponding GP is mean square differentiable (see e.g. [25] 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. [25] for a summary.

Proposition 3. ***Sample path differentiability.*** *With the same assumptions on $T_i(\cdot; \boldsymbol{\tau}_i)$'s and k_β as in proposition 2, and assuming in addition that*

- *D is compact,*
- *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 $\left.\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{(\boldsymbol{u},\boldsymbol{u})} + \left.\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{(\boldsymbol{u}',\boldsymbol{u}')} - 2\left.\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{(\boldsymbol{u},\boldsymbol{u}')} \leq \frac{C_0}{|\ln ||\boldsymbol{u}-\boldsymbol{u}'||^{1+\eta_0}}$,*

then the covariance c gives rise to a centred Gaussian Process possessing a version with differentiable sample paths.

Proof. 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. $C = C_0 \sum_{j=1}^{q} \sum_{j'=1}^{q} a_{j1} a_{j'1} \sup_{\boldsymbol{x} \in D} \left(T_j'\left(\boldsymbol{a}_1^\top \boldsymbol{x}\right)\right) \sup_{\boldsymbol{x} \in D} \left(T_{j'}'\left(\boldsymbol{a}_1^\top \boldsymbol{x}'\right)\right)$,
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 $\lim_{h \to 0} \left(\frac{\ln|\ln|h||}{\ln|\ln(C_T)+\ln|h||}(1+\eta_0) - 1\right) = \eta_0 > 0$).

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

$$
\begin{aligned}
&\left.\frac{\partial^2 c(\boldsymbol{u},\boldsymbol{u}')}{\partial u_1 \partial u'_1}\right|_{(\boldsymbol{x},\boldsymbol{x})} + \left.\frac{\partial^2 c(\boldsymbol{u},\boldsymbol{u}')}{\partial u_1 \partial u'_1}\right|_{(\boldsymbol{x}',\boldsymbol{x}')} - 2\left.\frac{\partial^2 c(\boldsymbol{u},\boldsymbol{u}')}{\partial u_1 \partial u'_1}\right|_{(\boldsymbol{x},\boldsymbol{x}')} \\
&= \sum_{j=1}^{q} \sum_{j'=1}^{q} a_{j1} a_{j'1} T_j'\left(\boldsymbol{a}_1^\top \boldsymbol{x}\right) T_{j'}'\left(\boldsymbol{a}_1^\top \boldsymbol{x}'\right) \left(\left.\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{\substack{(T(A\boldsymbol{x}'),\\ T(A\boldsymbol{x}'))}} + \left.\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{\substack{(T(A\boldsymbol{x}),\\ T(A\boldsymbol{x}))}} - 2\left.\frac{\partial^2 k_\beta(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{\substack{(T(A\boldsymbol{x}),\\ T(A\boldsymbol{x}'))}} \right) \\
&\leq \frac{C}{|\ln ||T(A\boldsymbol{x}) - T(A\boldsymbol{x}')||^{1+\eta_0}} \leq \frac{C}{|\ln ||\boldsymbol{x} - \boldsymbol{x}'||^{1+\eta}} \qquad (16)
\end{aligned}
$$

Using the theorem of sample path continuity for GP derivatives (see e.g. [41] 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. □

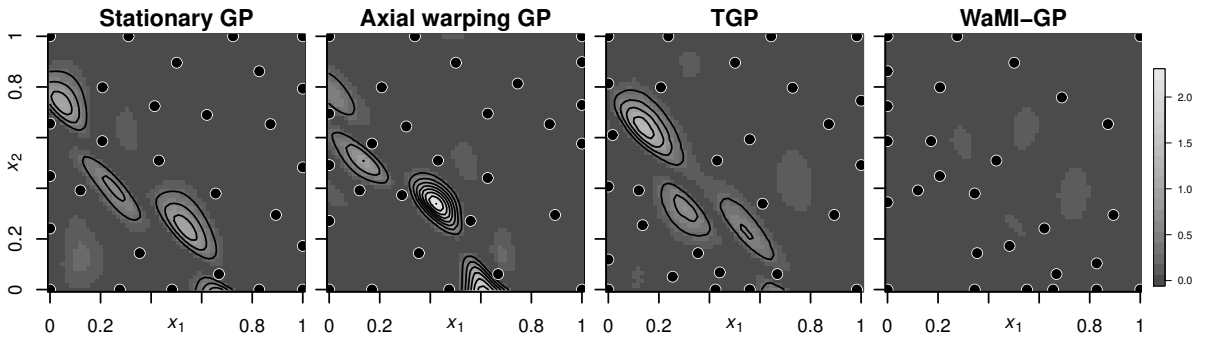

Figure 10: Prediction errors of four competing models on the running example function. The different models are a stationary anisotropic GP, an axial warping GP, Treed GP and WaMI-GP. For each method, we see the tenth step of a sequential design driven by the MSE criterion (shared initial design).

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

***Remark* 2. On the estimation procedure.** *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 already used earlier for illustrating a stationary model (Figure 2), the axial warping method (Figures 4), and the TGP model (Figure 5). 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 competitor models covered in the last section: stationary GP, GP with axial warping, and TGP. In Figure 10, we display the absolute difference between the real function and predictions from the four models after 10 sequential design steps based on the MSE criterion. Looking at the points selected 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 less evaluations in the flat regions (upper right). For the other models, prediction errors tend to occur in the high variation region. Hence our model, by detecting the high variation region and also associating a higher MSE there, enables to comparatively achieve enhanced prediction performance as illustrated in Figure 11.

These first results, that will be complemented by additional tests on two applications in Section 5, illustrate that WaMI-GP is able to account for heterogeneous regions in a semi-automated way (here all parameters including axes are estimated by MLE but the based kernel and the number of warping dimensions is fixed in advance) and hence to improve the performance of variance-based sequential design provided that some prior knowledge is available regarding the heterogeneities of the unknown function. In contrast, it might be the case that users do not feel confident to appeal to models where the number of parameters is inflated compared to the standard stationary situation, all the more so when the amount of available information on the objective function is drastically limited and the number of evaluations in the initial design is scarce. For those reasons we explore in the next section an alternative approach where the prior covariance is arbitrary (it may be a stationary one, a WaMI or any other kind of kernel) and the emphasis is put on infill sampling criteria rather than on the covariance structure. The goal is then to explore derivative-based criteria for the exploration of high-variation regions under

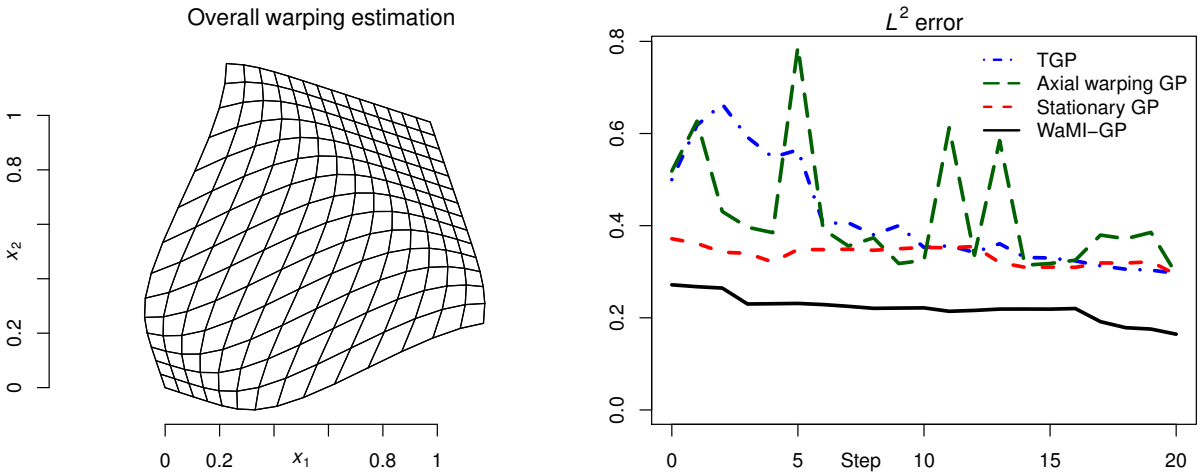

Figure 11: Prediction error of the running example function by the four considered models at each step of MSE-driven sequential designs.

any GP model. Later on is Section 5 the two approaches of working on kernels and/or on the sampling criteria for learning functions with heterogeneous variations will be compared and combined within a numerical benchmark. For now let us present and work out some novel infill sampling criteria based on GP derivatives.

4 Novel sampling criteria for detection of high variations

We saw in Section 3.4 above regarding the approximation of the running example function by diverse GP-related methods that exploring high-variation regions was key to quickly reduce the overall approximation error, hence explaining the good performances of WaMI-GP both in static conditions and in MSE-based sequential settings. Now, let us change the perspective by assuming that a prior covariance kernel is given (that may be thought of as a stationary kernel without any loss of generality) and putting the focus on infill sampling criteria dedicated to space exploration with an intensification on high-variation regions. With such a goal it is legitimate to aim at investing evaluation credit in regions where the data shows more local variability. A problem however with variance-based criteria such as considered so far in the paper is that they are homoscedastic in the observations, or in other words 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 Gaussian Process 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 [41]). Assuming indeed the differentiability

of m_n and the existence for all i of derivatives $\frac{\partial^2}{\partial t_i \partial t_i'} c_n(\boldsymbol{t}, \boldsymbol{t}')\Big|_{\boldsymbol{t}=\boldsymbol{t}'=\boldsymbol{x}}$ for all $\boldsymbol{x}, \boldsymbol{x}' \in D$,

$$\mathbb{E}\left(\nabla Y_{\boldsymbol{x}} \mid \mathcal{A}_n\right) = \nabla m_n(\boldsymbol{x}) \tag{17}$$

$$\operatorname{cov}\left(\nabla Y_{\boldsymbol{x}} | \mathcal{A}_n\right) = \left(\frac{\partial^2}{\partial t_i \partial t_j'} c_n(\boldsymbol{t}, \boldsymbol{t}')\Big|_{\boldsymbol{t}=\boldsymbol{x}, \boldsymbol{t}'=\boldsymbol{x}'} \right)_{i,j=1,\ldots,d}. \tag{18}$$

From there, there are a number of ways scalar indicators can be defined that quantify local variations and related uncertainties. In this work we chose to focus essentially on variance-based criteria for (exponentiated) gradient norms. To this means, let us consider the squared gradient norm process $(Q_{\boldsymbol{x}})_{\boldsymbol{x}\in D}$ defined by

$$Q_{\boldsymbol{x}} = ||\nabla Y_{\boldsymbol{x}}||^2_{\mathbb{R}^d} = \nabla Y_{\boldsymbol{x}}^\top \nabla Y_{\boldsymbol{x}}. \tag{19}$$

Although the squared gradient norm is obtained by applying a simple operation (taking the squared Euclidean norm) to a vector-valued Gaussian process, 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 up on such 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 and/or computed efficiently, leading to practical infill sampling criteria. Let us generalize indeed MSE and IMSE criteria to the (exponentiated) gradient norm.

Definition 2 (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*

$$J_n^{\mathrm{GNV}}(\boldsymbol{x}) = \operatorname{var}\left(||\nabla Y_{\boldsymbol{x}}|| \mid \mathcal{A}_n\right) = \operatorname{var}\left(\sqrt{Q_{\boldsymbol{x}}} \,\middle|\, \mathcal{A}_n\right) \tag{20}$$

J_n^{GNV} can be straightforwardly generalized by elevating the norm to some power $\eta > 0$, leading to

$$J_n^{\mathrm{GNV},\eta}(\boldsymbol{x}) = \operatorname{var}\left(||\nabla Y_{\boldsymbol{x}}||^\eta \mid \mathcal{A}_n\right) = \operatorname{var}\left(Q_{\boldsymbol{x}}^{\eta/2} \,\middle|\, \mathcal{A}_n\right). \tag{21}$$

Note that while the transformed norm loses its homogeneity, we abusively refer to this criterion as "GNV with exponent η" or "$\mathrm{GNV}(\eta)$*". $\mathrm{GNV}(1)$ is hence the previous* GNV*, and in what follows we will also pay a particular attention to* $\mathrm{GNV}(2)$*. Besides this, this class of criteria can also be generalized in the same way as* IMSE *generalizes* MSE *by integration, defining* IGNV *by*

$$J_n^{\mathrm{IGNV},\eta}(\boldsymbol{x}) = \int_{\boldsymbol{u}\in D} \mathbb{E}\left(\operatorname{var}\left(Q_{\boldsymbol{u}}^{\eta/2} \,\middle|\, \mathcal{A}_n, Y_{\boldsymbol{x}}\right) \middle|\, \mathcal{A}_n\right) \mathbf{d}\boldsymbol{u}. \tag{22}$$

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

Proposition 4. *Let $\boldsymbol{x} \in D$ and denote by $(\lambda_i(\boldsymbol{x}))_{1\le i\le d}$ the eigenvalues of $\nabla\otimes\nabla^\top c_n(\boldsymbol{x}, \boldsymbol{x})$. Then, the* $\mathrm{GNV}(2)$ *criterion can be written as follows:*

$$J_n^{\mathrm{GNV},\eta=2}(\boldsymbol{x}) = 4\ \nabla m_n(\boldsymbol{x})^\top \nabla\otimes\nabla^\top c_n(\boldsymbol{x}, \boldsymbol{x}) \nabla m_n(\boldsymbol{x}) + 2\sum_{i=1}^{d} \lambda_i(\boldsymbol{x})^2. \tag{23}$$

Furthermore, the $\mathrm{GNV}(1)$ *criterion can be expanded as follows:*

$$J_n^{\mathrm{GNV},\eta=1}(\boldsymbol{x}) = ||\nabla m_n(\boldsymbol{x})||^2 + \operatorname{tr}\left(\nabla\otimes\nabla^\top c_n(\boldsymbol{x}, \boldsymbol{x})\right) - \mathbb{E}\left(\sqrt{Q_{\boldsymbol{x}}} | \mathcal{A}_n\right)^2. \tag{24}$$

Finally, the corresponding integral criterion with $\eta = 1$ writes

$$
\begin{aligned}
J_n^{\mathrm{IGNV},\eta=1}(\boldsymbol{x}) = & \int_D \left(||\nabla m_n(\boldsymbol{u})||^2 + \frac{1}{c_n(\boldsymbol{x},\boldsymbol{x})} \kappa_n(\boldsymbol{u},\boldsymbol{x})^\top \kappa_n(\boldsymbol{u},\boldsymbol{x}) \right) \, \mathbf{d}\boldsymbol{u} \\
& + \int_D \left(\mathrm{tr}\left(\nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \right) - \mathbb{E}\left(\mathbb{E}\left(\sqrt{Q_{\boldsymbol{u}}} | \mathcal{A}_n, Y_{\boldsymbol{x}} \right)^2 \middle| \mathcal{A}_n \right) \right) \, \mathbf{d}\boldsymbol{u},
\end{aligned}
\tag{25}
$$

and its counter part in the case $\eta = 2$ can be expanded as

$$
\begin{aligned}
J_n^{\mathrm{IGNV},\eta=2}(\boldsymbol{x}) = & \int_{\boldsymbol{u}\in D} \left(4 \nabla m_n(\boldsymbol{u})^\top \nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \nabla m_n(\boldsymbol{u}) + 2 \sum_{i=1}^d \lambda_{i,\boldsymbol{x}}(\boldsymbol{u})^2 \right) \, \mathbf{d}\boldsymbol{u} \\
& + \frac{4}{\mathrm{var}\left(Y_{\boldsymbol{x}} | \mathcal{A}_n \right)} \int_{\boldsymbol{u}\in D} \kappa_n(\boldsymbol{u},\boldsymbol{x})^\top \nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \kappa_n(\boldsymbol{u},\boldsymbol{x}) \, \mathbf{d}\boldsymbol{u}
\end{aligned}
\tag{26}
$$

where $\lambda_{i,\boldsymbol{x}}(\boldsymbol{u})$ are the eigenvalues of $\nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \; = \mathrm{cov}\left(\nabla Y_{\boldsymbol{u}} | \mathcal{A}_n, Y_{\boldsymbol{x}} \right)$ 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 $\mathcal{A}_n$.

Proof of Proposition 4. 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:

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

The term $\mathrm{var}\left(\boldsymbol{Z}_{\boldsymbol{x}}^{c\top} \boldsymbol{Z}_{\boldsymbol{x}}^c \right)$ 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:

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

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

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

For the proof of $\mathrm{IGNV}_{\eta=1,2}$, we focus on the integrand. As for (23), we formulate the case $\eta = 2$ as follow:

$$
\begin{aligned}
\mathbb{E}\left(\mathrm{var}\left(Q_{\boldsymbol{u}} \, | \mathcal{A}_n, Y_{\boldsymbol{x}} \right) | \mathcal{A}_n \right) = & \, 4 \, \mathbb{E}\left(\mathbb{E}\left(\nabla Y_{\boldsymbol{u}} | \mathcal{A}_n, Y_{\boldsymbol{x}} \right)^\top \nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \mathbb{E}\left(\nabla Y_{\boldsymbol{u}} | \mathcal{A}_n, Y_{\boldsymbol{x}} \right) \middle| \mathcal{A}_n \right) \\
& + 2 \sum_{i=1}^d \lambda_{i,\boldsymbol{x}}^2.
\end{aligned}
\tag{27}
$$

We get the result with $\mathbb{E}\left(\nabla Y_{\boldsymbol{u}} | \mathcal{A}_n, Y_{\boldsymbol{x}}\right) = \nabla \boldsymbol{m}_n(\boldsymbol{u}) + \frac{Y_{\boldsymbol{x}} - m_n(\boldsymbol{x})}{c_n(\boldsymbol{x},\boldsymbol{x})} \kappa_n(\boldsymbol{u},\boldsymbol{x})$.
For $\eta = 1$, using the calculations for deriving (27), we obtain

$$\begin{aligned}\mathbb{E}\left(\operatorname{var}\left(\sqrt{Q_{\boldsymbol{u}}} \,\middle|\, \mathcal{A}_n, Y_{\boldsymbol{x}}\right) \middle| \mathcal{A}_n\right) =& \mathbb{E}\left(\left|\left|\mathbb{E}\left(\nabla Y_{\boldsymbol{u}} | \mathcal{A}_n, Y_{\boldsymbol{x}}\right)\right|\right|^2 \middle| \mathcal{A}_n\right) + \operatorname{tr}\left(\nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u})\right) \\ & - \mathbb{E}\left(\mathbb{E}\left(\sqrt{Q_{\boldsymbol{u}}} | \mathcal{A}_n, Y_{\boldsymbol{x}}\right)^2 \middle| \mathcal{A}_n\right). \end{aligned} \tag{28}$$

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

□

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

$$\mathbb{E}\left(\sqrt{Q_{\boldsymbol{x}}}\right) = \int_{\mathbb{R}} \sqrt{q} f_Q\left(q; \nabla m_n(\boldsymbol{x}), \nabla \otimes \nabla^\top c_n(\boldsymbol{x},\boldsymbol{x})\right) \mathrm{d}q \tag{29}$$

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

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 $c_n(\boldsymbol{x},\boldsymbol{x})$). Different methods for computing the distribution $f_Q(\cdot;\boldsymbol{\mu},\Gamma)$ of the quadratic form $Q = \boldsymbol{Z}^\top \boldsymbol{Z}$, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu},\Gamma)$, are summed-up, compared in [9] and implemented in a R package CompQuadForm. *The earliest method is based on approximating with a distribution of a* central *quadratic form, tuned for egalizing the three first moments [29] (equal skewness). The recent method of [22] is closely related to this, and provides error bounds. There exist also methods for which the error can be made arbitrarily small, that use numerical inversion of the characteristic function [20] or an infinite series formulation [12].*

Figure 12 displays the values of the four criteria, GNV and IGNV for $\eta = 1, 2$, in case of the running example test function (equation (3)) with the GP model of Figure 2. Integrated criteria (IMSE, IGNV) require more computational resource, 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, we see that 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, we notice that 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 figures suggest that integrated gradient-based criteria should be useful as expected in order to perform some kind of compromise between global uncertainty reduction and focus on high variations. This will be investigated in the next section, where the different approaches developed throughout the article will be tested and compared based on two engineering test cases.

5 Applications

This section deals with two numerical applications coming from a nuclear safety study conducted by IRSN 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, be it in terms of criteria (MSE, IMSE, $\mathrm{GNV}_{\eta=1,2}$, $\mathrm{IGNV}_{\eta=1,2}$) or surrogate

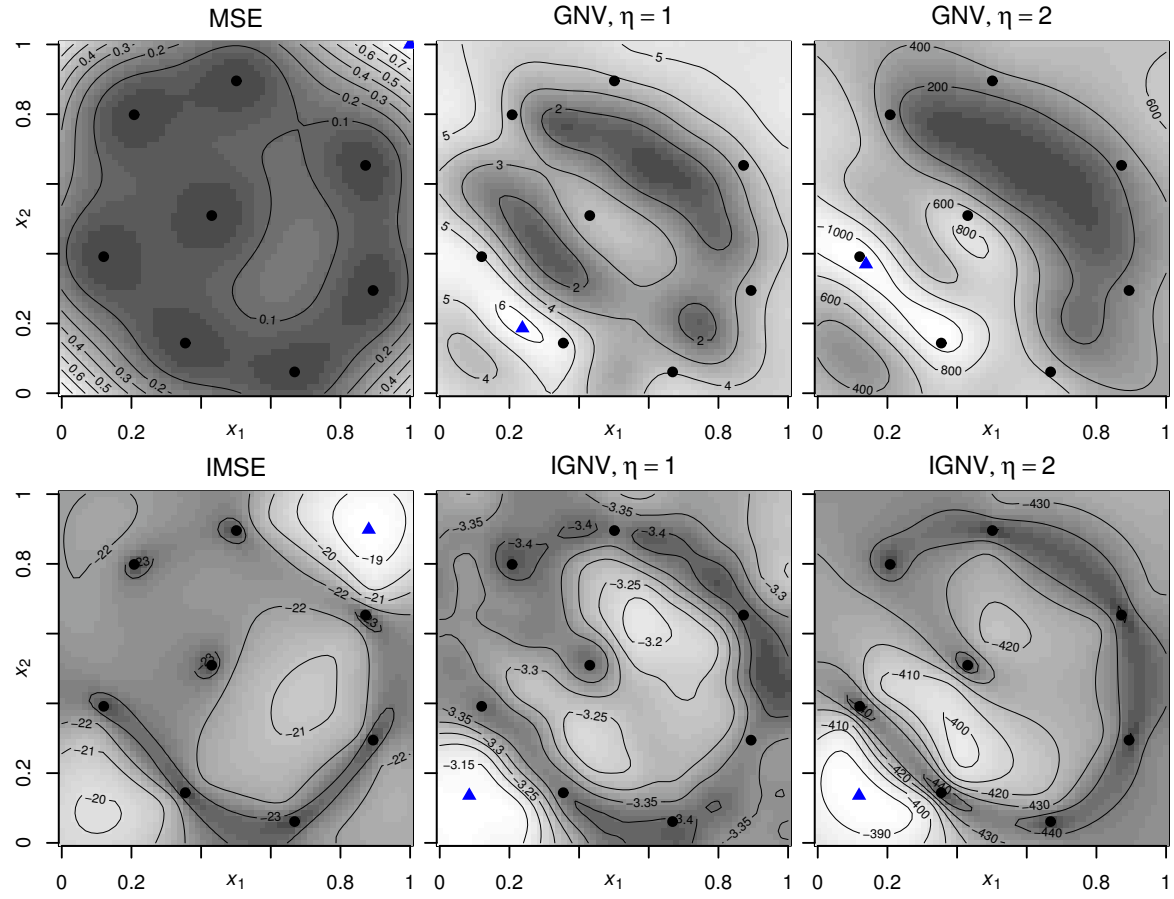

Figure 12: Classical and proposed criteria according to a stationary Gaussian process modelling.

models (stationary anisotropic, TGP, WaMI-GP). This assessment is achieved in both test cases by focusing on estimates of the L^2 prediction error:

$$\Delta = \sqrt{\int_D (\mu(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}}, \tag{30}$$

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 $\mathrm{IGNV}_{\eta=1,2}$ criteria, where the mean value of $Y_{\boldsymbol{x}}$ is plugged in the integrand, more precisely,

$$J_n^{\mathrm{IGNV}_{\mathrm{plugin}},\eta}(\boldsymbol{x}) = \int_{\boldsymbol{u}\in D} \mathrm{var}\left(||\nabla Y_{\boldsymbol{u}}||^{\eta} \,|\, \mathcal{A}_n, Y_{\boldsymbol{x}} = m_n(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{u}. \tag{31}$$

5.1 Cracking simulation of heterogeneous materials

This test case concerns mechanical studies in nuclear installations. More precisely, the objective is to analyse the crack propagation inside an heterogenous material such as concrete using the IRSN *Xper* code [30]. Two input variables are considered, related to geometrical and mechanical properties of the material. They are denoted W (ratio of interface energy) and L (inclusion length). The output of interest is the cracking energy which is the smallest energy required to break the material apart. Simulation times are long: depending on the input, they vary from

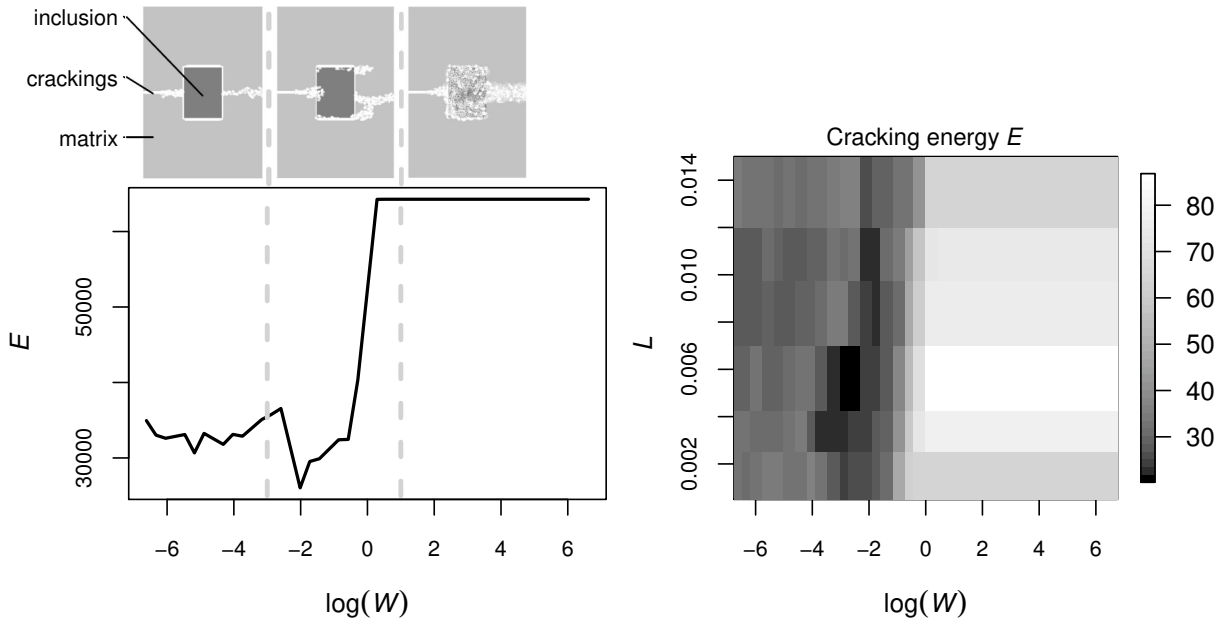

Figure 13: Cracking energy of an heterogeneous material depending on a mechanical parameter, $\log(W)$ the logarithm of the ratio of interface energy. The images (top left) represent three different crackings of a component. According to the inputs, we see that it propagates around (left) or through (right) the inclusion. These two modes correspond respectively to high or low cracking energy. A transition zone appears in between with high variations.

one day to one week. Therefore, evaluations should be chosen carefully in order to capture the function behaviour. The available dataset includes 216 points corresponding to the simulation of the response on a 36×6 grid (Figure 13). We show a 1D cut of the cracking energy with respect to $\log(W)$. We observe a "cliff", as it is sometimes called in the applied literature: a region where a small variation of the inputs impacts drastically the output. Coming back to Figure 13, this high variation zone is located along a straight line, slightly non-aligned to the canonical axes. Several series of tests are conducted. For 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. We built 5000 space-filling designs of size 20 (a set of random LHS designs 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. For the WaMI-GP covariance, we take for T_1 and T_2 cumulative distribution functions of beta distibutions. The results displayed on Figure 14 indicate that our approach outperforms the two other ones in terms of L^2 prediction errors.

It is also informative to analyse the estimated (overall) warpings, as displayed by Figure 15 (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 the 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 somehow 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.

A second test is performed to investigate the capability of the proposed criteria in *sequential* design of experiments procedures. For each criterion (MSE/IMSE, $\text{GNV}_{\eta=1,2}$/$\text{IGNV}_{\eta=1,2}$), we repeat 10 steps of the sequential design: point selections coupled with model updates. We choose

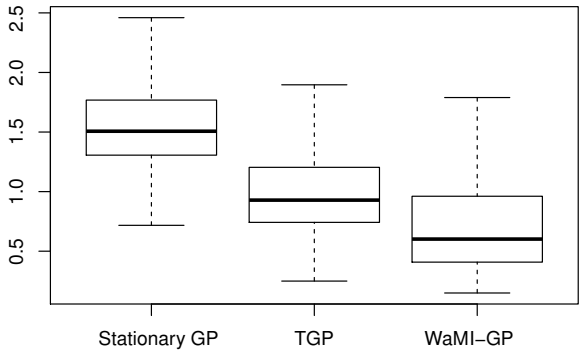

Figure 14: Comparison of L^2 prediction errors on the IRSN test case between the three candidate models: stationary anisotropic GP, TGP and WaMI-GP. The boxplots are obtained from repetitions with 5000 different initial designs (More detail in the text).

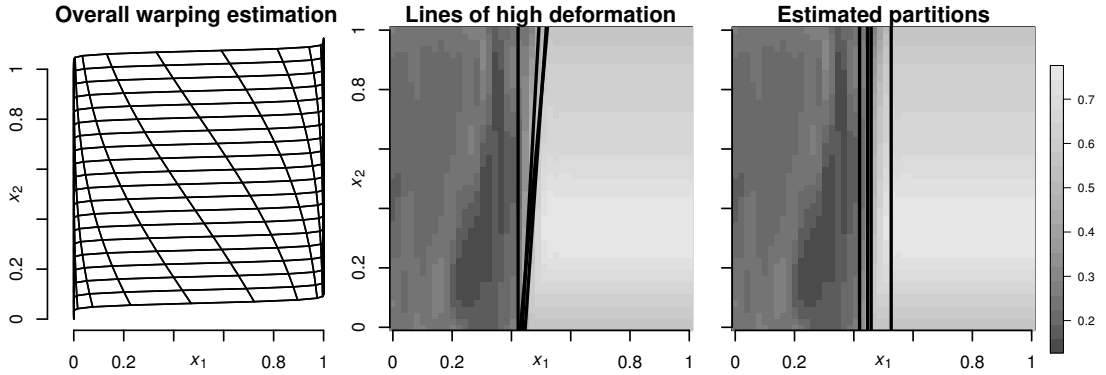

Figure 15: Some features of models with median prediction errors. Left: estimated warping of the WaMI-GP model with median predictivity; middle: lines of maximal distortion for 5 (most) median models; right: lines of partitioning for 5 median TGP models.

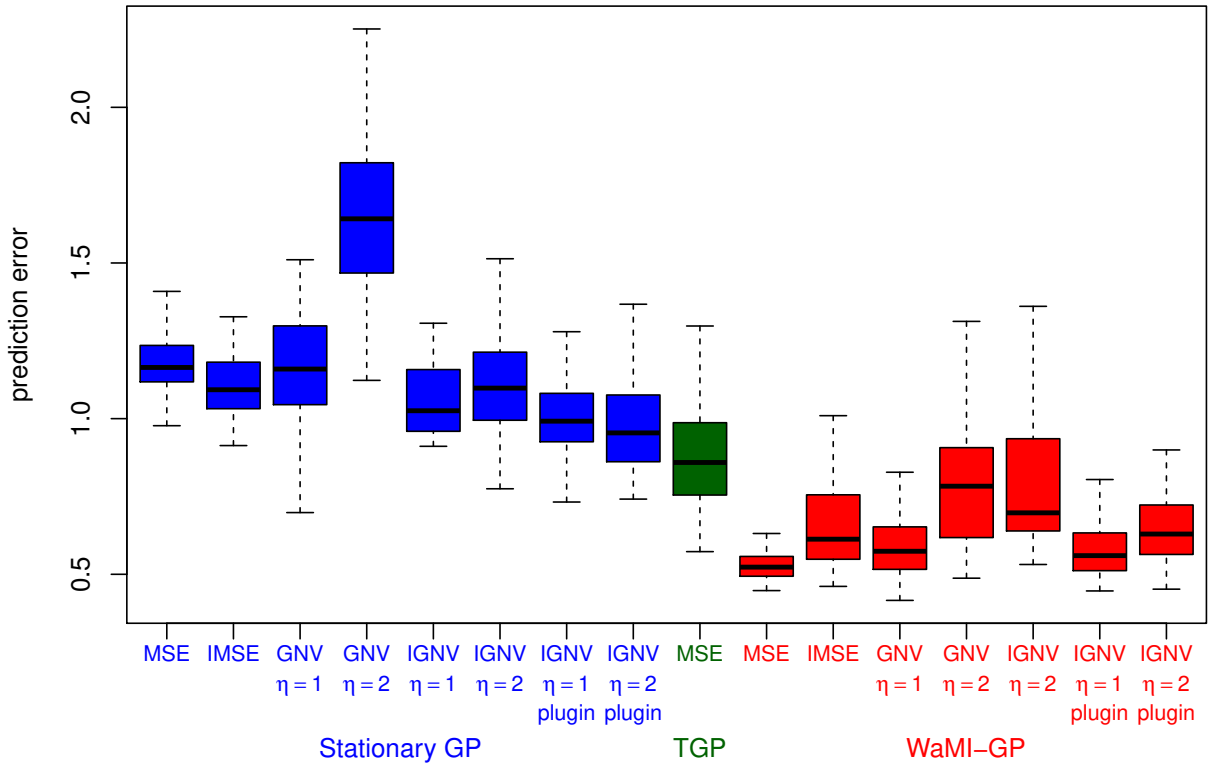

Figure 16: Distribution of the prediction error after different sequential design of experiments. Sampling criteria are compared in both stationary and our non-stationary models.

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 6 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

to take as GP models a stationary isotropic and a WaMI one. The whole workflow is replicated 100 times with a space-filling design of $n = 20$ points uniformly drawn among optimized LHS designs. The results are displayed in Figure 16 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 $\text{IGNV}_{\eta=1}$ 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 non-stationary and well-adapted to the behavior of f, the IMSE focuses naturally on the high variation zone and allows a reduction of about 30% of the number of simulations compared to the stationary framework. Finally, coupling WaMI-GP modelling and gradient-based criterion leads to rather poor results on Figure 16 since, by construction, both aspects contribute to exploitation with detrimental consequences on the exploration side.

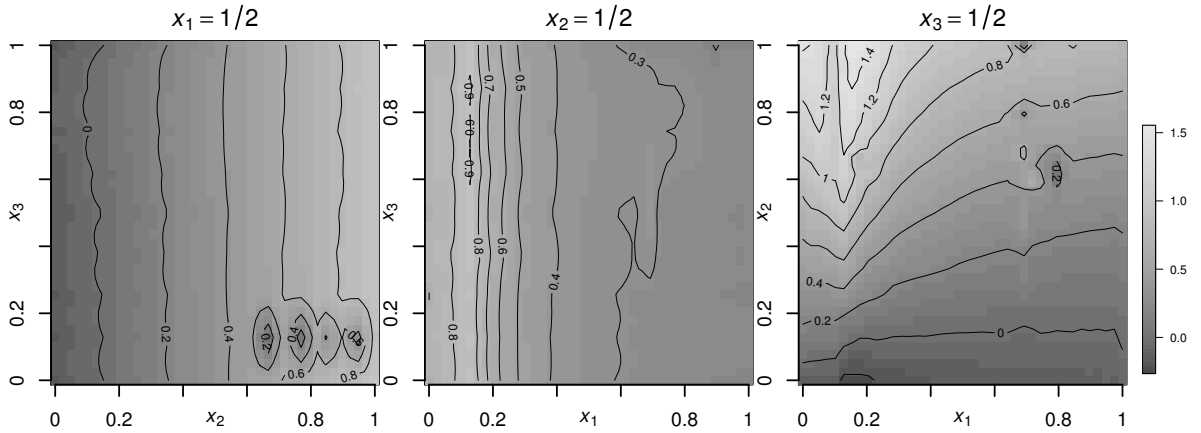

Figure 17: Simple three-dimensional interpolation of the available data on the Langley Glide-Back Booster simulation test case (only 3 slices of the input cube are displayed).

Table 2: Prediction errors of the models.

	$N_0 = 50$		$N_0 = 100$		$N_0 = 150$		$N_0 = 300$		$N_0 = 700$	
	median	$q_{95\%}$	median	$q_{95\%}$	median	$q_{95\%}$	median	$q_{95\%}$	median	$q_{95\%}$
WaMI-GP	6.887	7.889	6.516	7.057	6.326	6.765	6.135	6.206	6.750	6.825
TGP	6.879	7.980	6.095	7.424	5.798	7.647	5.569	6.409	4.954	5.576

5.2 The Langley Glide-Back Booster simulation

The Langley Glide-Back Booster is a rocket booster developed at NASA. Its behaviour is studied via numerical simulations. More details on the system behaviour and purpose are provided in [35]. Three input variables of the computer code controlling the trajectory of the rocket are considered: speed (measured in mach), angle of attack (alpha angle), and sidelip angle (beta). The output of interest is the lift force. Available data are displayed in Figure 17. We see that variations are this time mainly directed by canonical axis. The zone around the plane of equation $x_1 = 0.1$ (i.e. around mach one) has higher variations than in the rest of the domain, where the function is smoother. This calls for a non-stationary model. Some discontinuity is suggested by the data, which can be observed for example at the bottom right of the first plot (region $x_1 \approx 0.5, x_2 \geq 0.5, x_3 \leq 0.5$). These are due to the complexity of the simulator whose convergence depends on a solver which sometimes returns inaccurate values despite automatic checks [17]. Thus the models will be next considered in noisy settings in order to smooth out the convergence errors.

We investigate the capability of our non-stationnary covariance introduced in Definition 1 to model and predict the code output with and without sequential design of experiments. In order to study the influence of a fixed initial number of evaluations (maximin-optimized LHS), several tests are performed for designs including from 50 to 700 points. 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 median and 95% quantile of the errors. Table 2 provides the prediction error associated with our WaMI-GP approach vesus TGP [19].

It turns out that for a small training dataset, WaMI-GP leads to similar predictive performance as TGP in terms of median error but with a slightly lower 95%-quantile. When increasing the number of points in the initial design, TGP outperforms WaMI-GP. This makes sense as TGP model increases its complexity (i.e. its number of partitions and estimated parameters) accord-

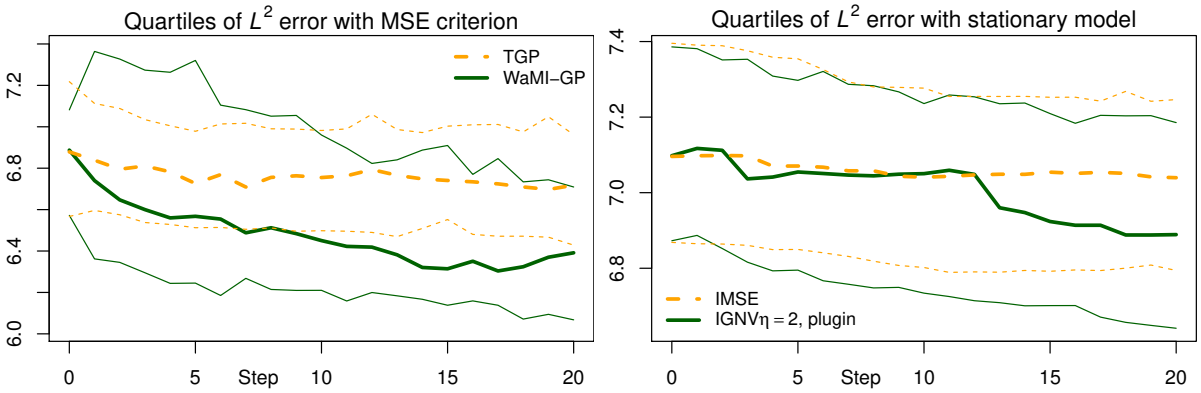

Figure 18: Medians (plain) and quartiles (dotted) of prediction errors during sequential designs of experiments. Left: comparison of models TGP and WaMI-GP with a common criterion MSE. Right: comparison of criterion IMSE and IGNV, $\eta = 2$, with a standard stationary GP model.

ing to the data while in its present form WaMI-GP has a fixed structure prescribed by the user. However, as we develop now, WaMI-GP outperformed TGP when adding experiments based in the MSE criterion.

Let us now focus indeed on sequential design of experiments. 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 to dilate the input space in the region of high variations (around mach 1). It results as an increased model variance in this area and thus a more dense 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], where similar variance-based criteria, Active Learning-MacKay and Active Learning-Cohn, are used in a pure Bayesian Framework). We also compared again a gradient-based criterion, IGNV, $\eta = 2$ with a classical criterion IMSE relying on a stationary anisotropic GP model (Figure 18).

We see that the IGNV($\eta = 2$) criterion leads to slightly lower prediction errors than 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 high variation region with the gradient-based criterion. To conclude this experimental section, reinforcing exploration in high-variation regions appears as a sound option to improving predictivity of surrogate models such as GPs, be it through adapted non-stationary covariances or via sampling criteria dedicated to this goal.

6 Conclusion

We have proposed non-stationary modelling and sampling approaches for predicting and designing experiments in a context of function with heterogeneous variations. The proposed WaMI-GP (Warped Multiple Index Gaussian Process) model was introduced, showing its link with existing modelling approaches such as Multiple Index modelling and the non-linear map method. We presented conditions under which the WaMI kernel is provably strictly positive definite and the corresponding centred GP is mean-square differentiable or respectively, possesses differentiable sample paths almost surely. We applied the model on toy examples and on functions from engineering case studies in dimensions 2 and 3. Although the number of parameters of WaMI-GP is kept affine rather than an exponential in the dimension, thanks to component-by-component univariate warpings, performances are competitive with respect to stationary GP and Treed

Gaussian Process (TGP) modelling. With bigger data sets, we experienced 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 where added by MSE maximization. It is also relevant to point that in case of a high variation zone slightly not aligned with a canonical axis, our model is favoured because its linear component can estimate an appropriate rotation of the data before the non-linear warping (note that non-axial partition in TGP is not currently implemented but is theoretically possible and could be implemented). In contrast, our method directly inherits from the non-linear 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 "stationarisation".

From a different viewpoint, we have also constructed novel criteria in sequential design of experiments for exploring function with high variation regions. These criteria are based on the gradient norm variance (GNV) of the modelling GP. These criteria 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 for adaptively approximating functions arising from the two engineering case studies. Numerical results are different according to the model. 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 trustful 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 non-stationarity 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 the Stepwise Uncertainty Reduction paradigm [2], weighted IMSE approaches [31] or other. Also, batch-sequential versions of the proposed criteria and their extensions owe to be defined and worked out. Focusing finally on the WaMI-GP model, several directions call for additional research. This includes notably investigations on efficient estimation algorithms for higher dimensions beyond brute force likelihood maximization, and also model selection versus full Bayesian approaches for inferring q and further parameters, pertaining for instance to the univariate deformations.

References

[1] E. B. Anderes and M. L. Stein. Estimating deformations of isotropic gaussian random fields on the plane. *The Annals of Statistics*, pages 719–741, 2008.

[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. *Statistics and Computing*, 22(3):773–793, 2011.

[3] D. Brillinger. The identification of a particular nonlinear time series system. *Biometrika*, 64:509–515, 1977.

[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 (4):455–465, 2014.

[5] C. Chevalier, D. Ginsbourger, and X. Emery. Corrected kriging update formulae for batch-

sequential data assimilation. In *Mathematics of Planet Earth*, pages 119–122. Springer, 2014.

[6] T. Choi, J. Q. Shi, and B. Wang. A gaussian process regression approach to a single-index model. *Journal of Nonparametric Statistics*, 23(1):21–36, 2011.

[7] Y. Deville, D. Ginsbourger, Contributors O. Roustant, N. Durrande, Maintainer Y. Deville, Depends Rcpp, Suggests DiceKriging, and LinkingTo Rcpp. Package 'kergp'. 2015.

[8] J. Doob. Stochastic processes. *New York: John Wiley & Sons*, 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. *Computational Statistics & Data Analysis*, 54(4):858–862, 2010.

[10] D. Dupuy, C. Helbert, and J. Franco. DiceDesign and DiceEval: Two R packages for design and analysis of computer experiments. *Journal of Statistical Software*, 65(11):1–38, 2015.

[11] N. Durrande, D. Ginsbourger, and O. Roustant. Additive covariance kernels for high-dimensional gaussian process modeling. In *Annales de la Faculté de Sciences de Toulouse*, volume 21, pages p–481, 2012.

[12] R.W. Farebrother. The distribution of a positive linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3):332–339, 1984.

[13] M. Gibbs. *Bayesian Gaussian processes for regression and classification.* PhD thesis, University of Cambridge, 1997.

[14] D. Ginsbourger, O. Roustant, and N. Durrande. On degeneracy and invariances of random fields paths with applications in gaussian process modelling. *Journal of statistical planning and inference*, 170:117–128, 2016.

[15] 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*, pages 315–330. Springer, 2016.

[16] R. B. Gramacy. *Bayesian treed Gaussian process models.* PhD thesis, University of California Santa Cruz, 2005.

[17] R. B. Gramacy and H. K. H. Lee. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[18] R. B. Gramacy and H. K. H. Lee. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–145, 2009.

[19] R. B. Gramacy and H. Lian. Gaussian process single-index models as emulators for computer experiments. *Technometrics*, 54(1):30–41, 2012.

[20] J.P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, pages 419–426, 1961.

[21] D. R. Jones, M. Schonlau, and J. William. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[22] H. Liu, Y. Tang, and H. H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, 2009.

[23] D. J. C. MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.

[24] J. Mockus. Application of Bayesian approach to numerical methods of global and stochastic optimization. *J. Global Optim.*, 4:347–365, 1994.

[25] C. Paciorek. *Nonstationary Gaussian Processes for Regression and Spatial Modelling*. PhD thesis, dissertation, Carnegie Mellon University, Department of Statistics, 2003.

[26] C. Paciorek and M. Schervish. Nonstationary covariance functions for gaussian process regression. *Advances in neural information processing systems*, 16:273–280, 2004.

[27] E. Padonou and O. Roustant. Polar gaussian processes and experimental designs in circular domains. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1014–1033, 2016.

[28] J.-S. Park and J. Baek. Efficient computation of maximum likelihood estimator in a spatial linear model with power exponential covariogram. *Computer & Geosciences*, 27(1):1–7, 2001.

[29] E. S. Pearson. Note on an approximation to the distribution of non-central χ 2. *Biometrika*, (46), 1959.

[30] 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. *European Journal of Computational Mechanics*, 19:389–417, 2010.

[31] V. Picheny, D. Ginsbourger, O. Roustant, R. T. Haftka, and N.-H. Kim. Adaptive designs of experiments for accurate approximation of target regions. *Journal of Mechanical Design*, 132(7), 2010.

[32] L. Pronzato and Werner G. Müller. Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3):681–701, 2011.

[33] P. Ranjan, D. Bingham, and G. Michailidis. Sequential experiment design for contour estimation from complex computer codes. *Technometrics*, 50(4):527–541, 2008.

[34] C. R. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[35] 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, 2003.

[36] J. Rougier. A representation theorem for stochastic processes with separable covariance functions, and its implications for emulation. arXiv:1702.05599 [math.ST].

[37] O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by Kriging-Based Metamodelling and Optimization. *Journal of Statistical Software*, 51 (1):1–55, 2012.

[38] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

[39] P. D. Sampson and P. Guttorp. Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.

[40] T. J. Santner, B. J. Williams, and W. Notz. *The design and analysis of computer experiments*. Springer Science & Business Media, 2003.

[41] M. Scheuerer. *A comparison of models and methods for spatial interpolation in statistics and numerical analysis*. PhD thesis, Georg-August-Universität Göttingen, 2009.

[42] J. Snoek, K. Swersky, R. S Zemel, and R. P. Adams. Input warping for bayesian optimization of non-stationary functions. In *ICML*, pages 1674–1682, 2014.

[43] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

[44] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

[45] E. Vazquez and J. Bect. Sequential search based on kriging: convergence analysis of some algorithms. *In: ISI - 58th World Statistics Congress of the International Statistical Institute (ISI'11), Dublin, Ireland*, 2011. arXiv preprint arXiv:1111.3866.

[46] B. J. Williams, T. J. Santner, and W. I. Notz. Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica*, 10:1133–1152, 2000.

[47] Yingcun Xia. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484):1631–1640, 2008.

[48] Y. Xiong, W. Chen, D. Apley, and X. Ding. A non-stationary covariance-based kriging method for metamodelling in engineering design. *International Journal for Numerical Methods in Engineering*, 71(6):733–756, 2007.