École Doctorale : Mathématiques et Informatiques de Marseille (ED184)

Institut de Mathématiques de Marseille

THÈSE DE DOCTORAT

pour obtenir le grade de

DOCTEUR de l'ÉCOLE CENTRALE de MARSEILLE

Discipline : Mathématiques

WARPING AND SAMPLING APPROACHES TO NON-STATIONARY GAUSSIAN PROCESS MODELLING

par

MARMIN Sébastien

Directeurs : GINSBOURGER David, LIANDRAT Jacques

Co-encadrant : BACCOU Jean

Thèse réalisée en cotutelle avec l'université de Berne et dans le cadre d'un contrat de formation par la recherche à l'IRSN.

Soutenue le 12 décembre 2017

devant le jury composé de

BACCOU Jean	Ingénieur de recherche à l'IRSN	Co-encadrant
FILIPPONE Maurizio	Maître de conférence à EURECOM	Examinateur
GINSBOURGER David	*Dozent* à l'université de Berne et *senior researcher* à Idiap	Directeur
GRAMACY Robert	*Professor at Virginia Polytechnic and State University*	Rapporteur
LE PENNEC Erwan	Professeur associé à l'École polytechnique	Rapporteur
LIANDRAT Jacques	Professeur des universités à l'École centrale de Marseille	Directeur
MARREL Amandine	Ingénieure de recherche au CEA	Examinatrice
POUET Christophe	Professeur des universités à l'École centrale de Marseille	Examinateur

À Caroline

Acknowledgement – *Remerciements*

Jean Baccou, David Ginsbourger et Jacques Liandrat forment une équipe d'encadrement bienveillante avec laquelle ce fût un plaisir de travailler. J'exprime ici ma reconnaissance envers Jacques Liandrat pour ses nombreux conseils. Je remercie également David Ginsbourger pour sa supervision enthousiaste et pour son important soutien, tant sur le plan scientifique qu'humain. Je remercie de même Jean Baccou pour l'encadrement déterminant qu'il m'a apporté.

I would like to thank Robert Gramacy and Erwan Le Pennec for having accepted to review the thesis, for their precise reports and detailed suggestions. Beside the reviewers, I am grateful to the jury members Maurizio Filippone, Amandine Marrel, and Christophe Pouet, for their constructive remarks.

Les enseignements reçus aux Mines de Saint-Étienne puis lors de mon stage à Berne m'ont servi pour ma thèse. J'en remercie particulièrement Xavier Bay, Éric Touboul, Olivier Roustant, Mireille Batton-Hubert, Rodolphe Le Riche, Nicolas Durande, Yann Richet, Victor Picheny, Clément Chevalier et David Ginsbourger.

During these three years, I had the opportunity to work in three different institutions. My sincere thanks also goes to Lutz Dümbgen, Ilya Molchanov and Johanna Ziegel, for welcoming me in their supportive team at the University of Bern. I thanks the members of the active IMSV, Dario, Mint, David, Clément, Tobias, Andrea, Ricardo, Lutz, Christof, Michael, Petro, Andrea, Florian, Qiyu...

Without the precious support of the *Institut de radioprotection et de sûreté nucléaire* it would not have been possible to conduct this work.

Je remercie tout les membres du LIMAR avec nos voisins de CS pour leur accueil amical à Cadarache, Gaétan, Nicolas, Jean, Sébastien, Éric, Olivier, Marc, Tony, Jimmy, Pierre, Christine, Tatiana, Daniele, Oussama, Valentin, Débora, Sébastien(s).

Je remercie également les nombreuses personnes avec qui j'ai eu l'opportunité d'échanger pendant la thèse, notamment aux rencontres ReDice et MASCOT-NUM, Luc Pronzato, Bertrand Iooss, Yann Richet, Jean Giorla, Esperan Padonou, Céline Helbert, Victor Picheny, Julien Bect, Hassan Maatouk, Nicolas Garland, Olivier Roustant, Nicolas Durrande, Guillaume Perrin, Bruno Torrésani ...

Je tiens maintenant à remercier l'équipe de Marseille, Jacques Liandrat, Sarah Mostefa, Thibaut Le Gouic, Chritophe Pouet, Renaud Bourlès, Harold Berjamin, avec qui j'ai eu le plaisir d'enseigner à Centrale.

Immense merci à ma femme Caroline, à mes parents Sabine et Gérard et ma marraine Muriel.

Contents

Notations

$\mathbb{N}^\star$	set of positive integers
$i, j, r, \ell, i', \dots$	mute index variables
$\#\boldsymbol{a}$	cardinality of vector or set $\boldsymbol{a}$
$\mathbb{R}$	set of real numbers
$]0, 1[$	open interval
$(\cdot)_+$	positive part of $(\cdot)$, i.e. $(x)_+ = \max(x, 0)$
$h, \varepsilon, \boldsymbol{h}$	real numbers and vector (often for infinitesimal limits)
$L^2(\mathbb{R})$	space of square integrable functions over $\mathbb{R}$
$\mathcal{C}^r(\mathbb{R}^d)$	set of real-valued function on $\mathbb{R}^d$ r times continuously differentiable
$\mathcal{C}^\infty(\mathbb{R})$	space of infinitely differentiable functions
$\mathcal{S}(\mathbb{R})$	Schwartz space, i.e. the space of functions C^∞ on $\mathbb{R}$ vanishing rapidly at infinity
S^p_{++}	set of positive definite matrices of size $p \times p$
$0_{\mathbb{R}^{\mathcal{D}}}$	null function on $\mathcal{D}$
$\det(B)$	determinant of a square matrix B
$\alpha_1, \dots, \alpha_N$	real numbers
$\lvert x \rvert$	absolute value of a $x \in \mathbb{R}$
$\mathbf{1}_p$	vector of ones in $\mathbb{R}^p$
$\boldsymbol{y} \to \mathbb{1}_{\mathcal{B}}(\boldsymbol{y})$	indicator function of $\mathcal{B}$, a subset of $\mathbb{R}^r$, $r \in \mathbb{N}^\star$
g, g_1, g_2, $\dots$	real-valued (multivariate) functions
$g_1 \circ g_2$	chaining of functions
$\bigotimes_{i=1}^r g_i$	tensor product of functions g_i's
$\boldsymbol{g}$	vector-valued function from $\mathcal{D}$ to $\mathbb{R}^p$
C, Σ, Σ', Γ, Γ', $\dots$	covariance matrices
$\varphi_{n,\Sigma}(\cdot)$	probability density of a centred n-variate normal distribution with non-singular covariance Σ
$\Phi_{q,\Sigma}$	q-variate cumulative normal distribution of covariance matrix Σ
f	objective function
$\mathcal{D}$	input set, typically $\mathcal{D} \subset \mathbb{R}^d$
$\boldsymbol{x}, \boldsymbol{x}'$	mute variables in $\mathcal{D}$
$x_{i,j}$	j^{th} coordinate of vector $\boldsymbol{x}_i$
n	a number of evaluations
N	total evaluation budget
n_0	size of initial design
q	batch size
X	a batch of q new points $(\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_{n+q}) \in \mathcal{D}^q$
$\check{X}$	a candidate batch of q points $(\check{\boldsymbol{x}}_{n+1}, \dots, \check{\boldsymbol{x}}_{n+q}) \in \mathcal{D}^q$
$X_{i_1:i_2}$	shortcut notation for $\left(\boldsymbol{x}_{i_1}^\top, \boldsymbol{x}_{i_1+1}^\top, \dots, \boldsymbol{x}_{i_2}^\top\right)^\top$ for i_1, $i_2 \in \mathbb{N}$, $i_1 \leqslant i_2$
$(\Omega, \mathcal{F}, \mathbb{P})$	probability triplet
$\mathcal{B}$	Borel σ-algebra
$L^2(\Omega, \mathbb{P})$	space of random variables with finite variance
U, V	random variables
$\boldsymbol{X}$	a random vector in $\mathcal{D}$
$\mathcal{GP}(m, c)$	Gaussian process (GP) distribution with mean m and covariance c
Y, Z	Gaussian processes
m, c	Y's mean and covariance functions
μ, k	Z's mean and covariance functions

$\boldsymbol{m}$	vector of evaluations of m at $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
$\mathcal{A}_n$	evaluation event $\{Y_{\boldsymbol{x}_1} = y_1, \ldots, Y_{\boldsymbol{x}_n} = y_n\}$, $y_1, \ldots, y_n \in \mathbb{R}$
C	covariance matrix associated with c at points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$: $C = (c(\boldsymbol{x}_i, \boldsymbol{x}_j))_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$
$\boldsymbol{Y}_{1:n}$	Gaussian vector $(Y_{\boldsymbol{x}_1}, \ldots, Y_{\boldsymbol{x}_n})^\top$
$\boldsymbol{Y}, \boldsymbol{Y}^{(a)}, \boldsymbol{Z}, \ldots$	Gaussian vectors
$\boldsymbol{m}^{(a)}, \boldsymbol{m}^{(b)}$	$\boldsymbol{Y}^{(a)}$ and $\boldsymbol{Y}^{(m)}$ mean vectors
$C_{a,a}, C_{b,b}, C_{a,b}$	covariance and cross-covariance matrices for $\boldsymbol{Y}^{(a)}$ and $\boldsymbol{Y}^{(b)}$
$\boldsymbol{m}_{b \mid a}, C_{b \mid a}$	conditioned mean and covariance (proposition 1)
$\boldsymbol{c}_n$	vector-valued function $\boldsymbol{x} \to (c(\boldsymbol{x}, \boldsymbol{x}_i))_{i=1,\ldots,n}^\top$
m_n, c_n	posterior mean and covariance of Y knowing $\mathcal{A}_n$
σ	prior standard deviations, $\boldsymbol{x} \to \sqrt{c(\boldsymbol{x}, \boldsymbol{x})}$
σ_n	posterior standard deviations $\boldsymbol{x} \to \sqrt{c_n(\boldsymbol{x}, \boldsymbol{x})}$
$k^{\text{`}5/2\text{'}}$	(univariate) Matérn kernel with smoothness parameter $\nu = 5/2$
$k^{\text{`}\infty\text{'}}$	Gaussian kernel
ν	Matérn smoothness parameter
θ	correlation length (for univariate GP)
$\widehat{\beta}$	estimated constant trend in ordinary kriging
$\widehat{\boldsymbol{\beta}}$	estimated vector of trend coefficients in universal kriging
$\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$	parameter vector and its estimator
$\boldsymbol{\beta}$	unknown trend coefficients in universal Kriging
G	$n \times p$ matrix $(\boldsymbol{g}(\boldsymbol{x}_1), \ldots, \boldsymbol{g}(\boldsymbol{x}_n))^\top$ in universal Kriging
$J_{n,q}^{\text{NAME}}$	a criterion defined with $\mathcal{A}_n$ for batchsize q
α	exponent of the improvement in the generalized q-EI
η	exponent parameter of GNV and IGNV criteria
$\Delta(\boldsymbol{\theta})$	quadratic prediction error of model with parameters $\boldsymbol{\theta}$
$\mathcal{I}$	a function from index set $\{1, \ldots, n\}$ to $\{1, \ldots, k\}$, $k \in \mathbb{N}^\star$
$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{1:n})$	for given parameters $\boldsymbol{\theta}$, likelihood of the evaluation values $\boldsymbol{y}_{1:n}$
$m_{\boldsymbol{\theta}_1}, c_{\boldsymbol{\theta}_2}$	mean and covariance parametrized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$
H_0	null hypothesis
Λ	likelihood ratio
Θ	a parameter space
p	number of parameters (or in another context, size of truncated Gaussian vectors in q-EI formula)
$m_{n,\boldsymbol{\theta}}^{(i)}(\boldsymbol{x}_i)$	prediction at point i without the evaluations of the group $\mathcal{I}(i)$
$(\Sigma_{\boldsymbol{x}})$	family of covariance matrices for convolution method
$\mathcal{E}$	latent warped space
dist_Γ	Mahalanobis distance with symmetric definite matrix Γ
$\boldsymbol{a}, A$	vector and matrix, parameter for SIM-GP and MIM-GP
$\boldsymbol{b}$	parameter vector of $k_{\boldsymbol{b}}$
d'	Number of axial warpings in WaMI-GP Model
$p_1, \ldots, p_{d'}$	size of parameter vectors (for the axial warpings for the WaMI-GP model)
$\boldsymbol{\gamma}$	warping of the input space $\mathcal{D}$
γ	univariate or axial warping
$\boldsymbol{\rho}, \rho_i$	covariance parameters related to warpings
$\left.\frac{\partial g(\boldsymbol{t})}{\partial t_i}\right\vert_{\boldsymbol{t}=\boldsymbol{x}}, \frac{\partial g(\boldsymbol{x})}{\partial x_i}$	partial derivatives of a function g
$\nabla_{\boldsymbol{t}} \left[g(\boldsymbol{t})\right](\boldsymbol{x}), \nabla g(\boldsymbol{x})$	gradient of a function g
$\left.\frac{\partial^2 c(\boldsymbol{t},\boldsymbol{t}')}{\partial t_i \partial t'_{i'}}\right\vert_{\substack{\boldsymbol{t}=\boldsymbol{x} \\ \boldsymbol{t}'=\boldsymbol{x}'}}$	cross derivatives of a covariance function c
$\frac{\partial^2 c}{\partial x_i \partial x'_{i'}}$	the function $(\boldsymbol{x}, \boldsymbol{x}') \to \left.\frac{\partial^2 c(\boldsymbol{t},\boldsymbol{t}')}{\partial t_i \partial t'_{i'}}\right\vert_{\substack{\boldsymbol{t}=\boldsymbol{x} \\ \boldsymbol{t}'=\boldsymbol{x}'}}$ on $\mathcal{D}^2$
g'	derivative of a univariate function g
$\frac{\partial Y_{\boldsymbol{x}}}{\partial x_i}, \nabla Y_{\boldsymbol{x}}$	partial derivative or gradient of Y (by mean-square or sample path differentiability)
$C, {}^{\nabla}C, {}^{\nabla'}C, {}^{\nabla^2}C$	block matrices forming the covariance of $[Y_{\boldsymbol{x}_{n_0+1}}, \ldots, Y_{\boldsymbol{x}_{n_0+q}}, \nabla Y_{\boldsymbol{x}}]$
$\mathcal{W}_y$	wavelet transform of a signal y
τ, s	position and scale parameter of the wavelet transform

Chapter 1

Introduction

Context and motivations

In risk analysis of a complex system, it is crucial to ensure that moderate variations of input parameters will not move the system towards very different conditions from reference ones. A particularly challenging situation is when the sensitivity of the system to input perturbations substantially varies across the input space. Many systems abruptly change regime: in material science with percolation of porous media; in epidemiology with outbreak of a pathogen according to uncertain characteristics of a population; in signal and image processing with discontinuities or edges (i.e. abrupt changes in the grey levels) deteriorating the efficiency of compression algorithms; in thermodynamics with phase transition, and 'cliff effects' in mechanics where competing phenomena can generate steep transitions and strong gradients in the response of a material. This last contextual example is encountered in safety studies of composite materials used in nuclear plants conducted by the French Institute for Radiological Protection and Nuclear Safety (*Institut de Radioprotection et de Sûreté Nucléaire*, abbreviated IRSN).

Let us assume that we aim to study one real-valued response of some deterministic system with respect to d variables, formally an objective function $f : \boldsymbol{x} \in \mathcal{D} \subset \mathbb{R}^d \to f(\boldsymbol{x}) \in \mathbb{R}$. Abrupt changes of regime are reflected for instance, for differentiable f's, by changing magnitude of the gradient norm depending on regions of the input space or, to take an alternative viewpoint, by spatially-varying main local frequencies. Here we will informally refer to f possessing such features as "functions with heterogeneous variations".

Figures 1.1 to 1.3 show synthetic and real-world examples of functions with

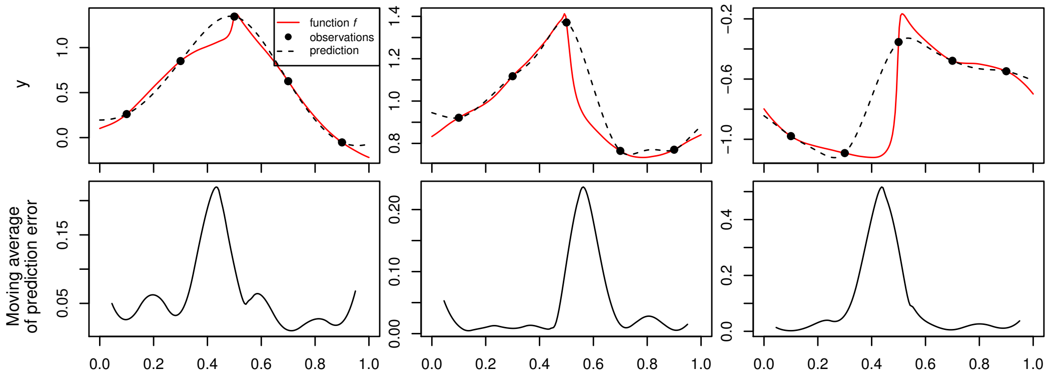


Figure 1.1: Top: functions with high variation zones, obtained by generating sample paths of a non-stationary warped Gaussian process (see section 2.2 for details), with predictors based on five evaluations. A classical interpolating model is used, more precisely a Gaussian process model with stationary covariance of type Matérn $\nu = \frac{5}{2}$, see section 2.1. Bottom: concentration of the prediction error around high variation zones (empirical assessment, displaying the absolute differences between predictions and true values, averaged on a moving window of width $\frac{1}{10}$).

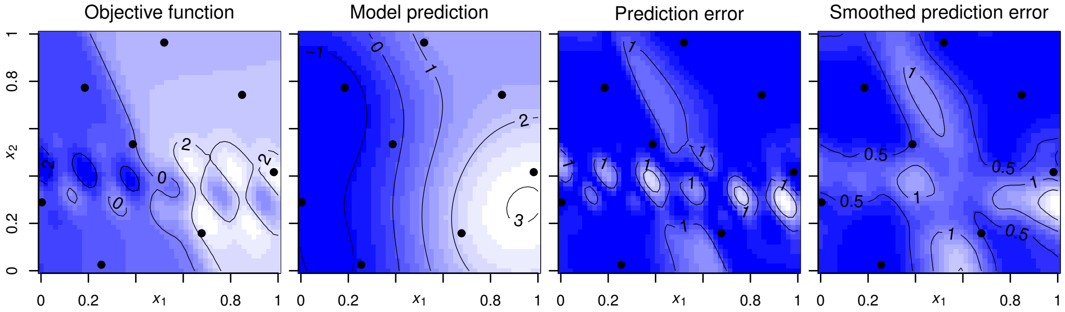


Figure 1.2: A function with heterogeneous variations (eq. (1.1)), its prediction from a standard model (stationary GP model, with covariance of type Matérn $\nu = 5/2$, isotropic, see section 2.1), model error, and its moving average of radius 0.01.

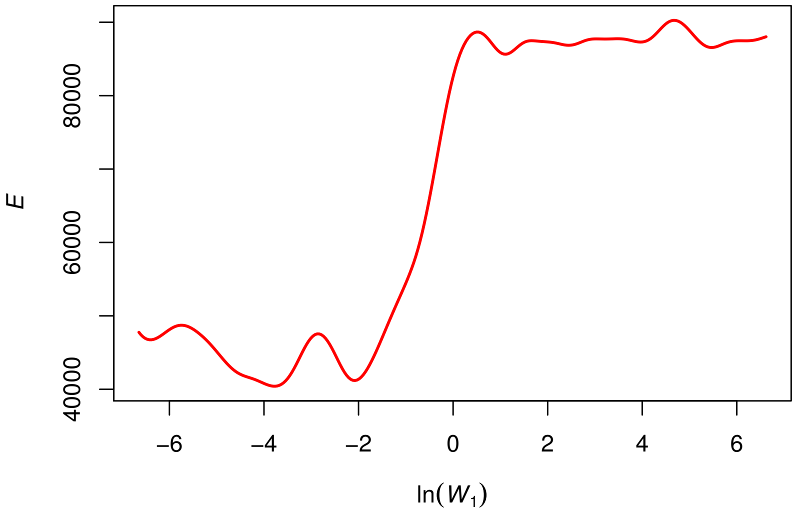


Figure 1.3: Example of an output of interest (cracking energy) with respect to a mechanical input controlling the cracking of a composite material subjected to a traction force.

heterogeneous variations in uni- and bivariate cases. The first univariate functions (fig. 1.1), drawn randomly[1], have a steep transition in the middle of the input domain. The bivariate toy function is defined for $(x_1, x_2) \in [0,1]^2$ by

$$\begin{aligned} f(x_1, x_2) = &\sin(2.5x_1)\cos(2x_2) + \arctan(30((x_1 + 0.5x_2) - 0.7)) \\ &+ (\sin(35(x_1 + 0.3x_2)) + \cos(20(x_2 + 0.2x_1)))e^{-50(x_2+0.1x_1-0.4)^2}. \end{aligned} \tag{1.1}$$

The one-dimensional function displayed in fig. 1.3 is extracted from fracture dynamics calculations for composite materials, arising from risk studies at IRSN. Evaluating this function can be considered as expensive – one to three days of computation for a single point – because it requires heavy high-fidelity numerical simulations.

Such expensive functions are typically encountered in the resolution of partial differential equations from physical sciences and engineering [Forrester et al., 2008], and more generally in all application fields that appeal to computer experiments [O'Hagan, 1978, Sacks et al., 1989, Schonlau, 1997]. The lack of data due to prohibitive evaluation costs makes the analysis of such functions challenging. One way around this problem is to rely on predictions of f. However, in our context of interest, using standard prediction approaches can lead

[1] as realisations of a warped Gaussian process, see later section 2.1.1

to poor predictions not only in scarcely explored region of the input space, but also in regions of high variations. To alleviate this issue, it is natural to think of reinforcing exploration in such areas. From a different perspective, one can also attempt to incorporate prior knowledge pertaining to the heterogeneous variations of f within the models used for prediction. Our main focus here is on Gaussian process (GP) models that have become quite popular in the last decades for approximating and exploring systems based on scarce evaluations and have become a standard in the design and analysis of computer experiments (see e.g. [Sacks et al., 1989, Jones et al., 1998, Santner et al., 2003]). Gaussian process models consist of assuming that the unknown objective function f is a sample path of a Gaussian process Y indexed by the input space of f. The function f is approached by a so-called 'posterior' stochastic process that takes the evaluations into account. This statistical (and Bayesian) framework offers efficient tools, notably for designing parsimonious and optimal evaluation strategies. Through a choice of a mean and a covariance functions, GP models are versatile and can integrate practitioner's initial knowledge on f. A good specification of the GP model as well as the use of adapted *sampling criteria*, i.e. predefined functions determining the next evaluations based on the current GP model, are two crucial aspects in GP-driven sequential design of experiments. We now focus on the first point.

First angle: modelling approaches

Adapting the covariance of Y to specific classes of objective functions f has inspired a lot of research. For example, for objective functions with a better representation in polar coordinate, [Padonou and Roustant, 2016] propose GP models that incorporate the geometry of the disk. Similarly, appropriate covariances exist for functions known to satisfy degeneracies such as symmetries or harmonicity [Ginsbourger et al., 2016a], and for functions with a sparse ANOVA decomposition [Durrande et al., 2012, Ginsbourger et al., 2016b]. In the absence of such specific assumption on f, it is common to take stationary covariance functions [Stein, 1999]. A stationary covariance is invariant by translation: for a GP with a constant mean, it implies that 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 $\boldsymbol{x} - \boldsymbol{x}'$.

Let us focus again on figs. 1.1 and 1.2 and discuss how standard GP models[2] actually perform when predicting two synthetic test functions that possess abrupt

[2] i.e. with constant mean and stationary covariance with here type Matérn and smoothness parameter $\nu = 5/2$, see section 2.1

changes in function values or local frequency. In both figures, we represent the absolute difference between objective functions and their predictions. These prediction errors are spatially averaged on short windows, localising where the model is less accurate. We observe in fig. 1.1 that the prediction errors are higher where f varies faster, i.e. it has higher derivatives. Similarly in fig. 1.2, we see more prediction errors in the high variation regions, localised around the line of equation $2x_1 + x_2 = 7/150$ (for the cliff) and around the line of equation $x_1 + 10x_2 = 4$ (for the change of average frequency). Although practitioners can often tell if such area of heterogeneity exists, they may possess only limited information on their locations, shapes or orientations.

When f is known to possess heterogeneous variations, it is sensible to depart from the stationary hypothesis and consider non-stationary covariances that account for this property. An appropriate non-stationary covariance can adapt progressively to the heterogeneous behaviour as long as it is re-estimated step-by-step as new evaluations become available. Among various proposals from non-stationary GP modelling, we consider later in section 2.2 convolution methods, see [Paciorek and Schervish, 2004, Gibbs, 1997], or input space warping approaches [Sampson and Guttorp, 1992]. We focus in this thesis on the latter, where non-stationary GPs come from the chaining of a GP with a warping of the space $\mathcal{D}$. A review of existing space warping approaches is in section 2.2. The main challenge here is to estimate the warping with arbitrary d-dimensional $\mathcal{D}$ from scarce evaluations. An important question to address in GP modelling, and in particular when using input space warping, is the balance between the flexibility and the sparsity of a model, i.e. the compromise between the capability to predict accurately diverse types of functions, and having a low number of model parameters, for easing (or just enabling) the model estimation. In the present context of small data sets and heterogeneous variations, maintaining sparsity of GP models while keeping nice flexibility properties is a requirement.

The question of non-stationary modelling is also addressed in the field of signal or image processing. Capturing local variations in a signal or an image, seen as heterogeneity of a function, is possible with the well-known wavelet transform [Daubechies, 1992, Mallat, 1998]. It can detect breakdowns, and contrary to the Fourier transform, it informs about precise locations of the breakdowns as well as their scale levels (or ‘local frequencies’). The main challenge is here to adapt existing wavelet approaches to the context of computer experiments where regular and dense grids of evaluations are not available.

Second angle: sampling approaches

As an alternative to non-stationary covariance, allocating more evaluations in specific regions (for our purpose, regions of high variations) can be achieved by a sampling strategy based on a GP model. Sampling criteria are real-valued functions on $\mathcal{D}$ (or $\mathcal{D}^q$ in the case of batch-sequential design) scoring the relevance of evaluating next at any candidate point $\boldsymbol{x} \in \mathcal{D}$. Maximisation of the criteria and updates of the model with new evaluations are then repeated until a stopping condition is met, e.g. depletion of the evaluation budget. Criteria classically derive from the posterior variance for allocating evaluations to unexplored regions as for example the Mean Squared Error (MSE) and Integrated MSE (IMSE) criteria [Sacks et al., 1989], detailed in section 2.1.2. While these strategies may eventually learn high-variation regions of a function with a heterogeneous behaviour, in stationary cases it is done in a non-adaptive way as the prediction covariance does not depend directly[3] on evaluation outputs but solely on the location of evaluation points. Therefore, to outperform these general purpose criteria in cases when a specific goal is predetermined (here take advantage of the capabilities of the model in a context of high heterogeneity and expensive evaluations), many specific criteria are developed for targeting the evaluation locations in areas of interest.

When the goal is to optimise f, a number of sampling criteria have been proposed in the literature [Jones, 2001, Frazier et al., 2008, Contal et al., 2014]. The *Expected Improvement* (EI) criterion [Mockus, 1989, Jones et al., 1998] and its multipoint version for batch evaluations are particularly popular in the literature for their intuitive definitions, and their properties (e.g. the one-step lookahead optimality [Ginsbourger and Le Riche, 2010]). Defined on $\mathcal{D}^q$, a multipoint criterion provides after maximisation a batch of points deemed most promising for parallel evaluations of f. This allows to distribute evaluations over several experimental units, as parallel computing became popular in recent years due to the fast development of clouds, clusters and GPUs.

In addition, several criteria are dedicated to further objectives, such as inversion, estimation of excursion set of f above a given threshold, probability of failure, etc. Associated design strategies aim at getting precise predictions of f in specific regions of interest, for instance around a contour lines of f with a threshold value. As examples we mention the targeted IMSE [Picheny et al., 2010], the Expected Improvement of Ranjan et al. [2008] which uses the difference between the posterior GP and the threshold, methods of Bect et al.

[3] The prediction covariance indirectly depends on the output values via the model estimation, see section 2.1.

[2011], Chevalier et al. [2014a] that focus on reducing the uncertainty on a volume of the excursion set, or approaches of Chevalier et al. [2013], Azzimonti [2016] where the emphasis is put on estimating excursion sets by exploiting notions from random sets theory. Several of the criteria of interest may be formulated within the framework of *stepwise uncertainty reduction* (SUR), see [Bect et al., 2017] and references therein. This framework aims at finding an optimal sequence of evaluation points in order to reduce a targeted uncertainty quantity.

Gramacy and Lee [2009] adapt variance-based sampling criteria for favouring high variation regions via a non-stationary model and facilitate multipoint asynchronous designs with unknown batchsize[4]. A natural idea that we will pursue here to build criteria favouring high variation regions is to exploit conditional distributions of partial derivatives of Y. In particular, the fact that these conditional distributions are Gaussian and with known moments will be a key to obtain tractable formulas for candidate sampling criteria.

Structure of the thesis

Chapter 2 is a state-of-the-art on GP models and design of experiments, with two main methodological foci: on non-stationary models and their use for sequential design of experiments and on EI sampling for global optimisation.

We then tackle contributions to modelling and sampling heterogeneous functions from two angles. For the first angle, in chapter 3 we rely on a new family of non-stationary covariances (WaMI, for warped multiple index) that simultaneously generalises features from Multiple Index GPs and tensorised warpings (presented in chapter 2). A GP model using with a WaMI covariance (WaMI-GP) is investigated through mathematical analysis. In particular, we explore its ability to approximate a quite wide family of functions while remaining tractable (with a moderate number of parameters to be inferred). Indeed, it is shown that the model can incorporate any orientation of heterogeneous variations, and besides this, the number of covariance parameters increases affinely with slope 1 with respect to the number of inputs. Also, independently of the WaMi covariance, an algorithm building a warped GP model, called Wav-GP, is proposed. It uses the local scale of a wavelet transform for the warping estimation.

For the second angle, chapter 4 is dedicated to sampling criteria. We study

[4] 'Asynchronous' means that the differences in evaluation times are taken into account in the construction of the design for improving the use of parallel computers.

several proposals of derivative-based criteria built for the exploration of high variation regions. These criteria are meant to make a trade-off between filling the space uniformly for a global exploration and intensifying the sampling in high variation regions for a faster reduction of prediction error. We conduct derivations of these criteria which rely on the variance of the GP gradient norm field, facilitating their optimisation. Our aim in this chapter is also to present a set of novel results pertaining to the calculation, the computation and the optimisation of the multipoint EI criterion. As most of these novel results apply to a broader class of criteria, we present a generalisation of the multipoint EI[5]. The obtained formula is then revisited, leading to a numerical approximation of the multipoint EI with arbitrary precision and very significantly reduced computation time. Moreover, approaches for fast gradient approximations with controllable accuracy are presented.

Chapter 5 deals with a series of applications of the methods developed in the previous chapters with a special attention to the comparisons with existing approaches. After describing the respective contexts of applications, the first part of the chapter has two aims. The first one is to show the expressiveness encoded by WaMI-GP models and to compare then with other modelling methods introduced in the next chapter. The second aim is to evaluate the performance of the derivative-based criteria compared to classical MSE and IMSE criteria in both stationary and non-stationary settings. These comparisons are made in particular on two mechanical test cases. The first test case stems from numerical simulations of fracture dynamics arising in risk studies at IRSN in the framework of the MIST lab activity. The second test case is a three-dimensional fluid dynamics application from NASA that was used in an article about the Treed Gaussian Process model [Gramacy and Lee, 2009].

Three further numerical experiments are performed. The first one is related to the implementation of a batch-sequential approach to function approximation under WaMI-GP modelling and its comparison with a baseline method, illustrating substantial speed-ups that can be of particular interest for industrial and further real-world applications. The second application concerns the application of the Wav-GP approach and highlights its potential for prediction of heterogeneous functions. The last application illustrates the accuracy of the proposed fast approximation of multipoint EI and its gradient, and the associated speed-ups obtained in multipoint EI maximisation experiments.

[5] Generalisation that allows accounting for noise in conditioning observations and also exponentiating the improvement.

Chapter 2

State-of-the-art of Gaussian process modelling and design of experiments, with a focus on non-stationarity

We focus on Gaussian process (GP) models, popular for approximating and exploring non-linear systems based on a small number of evaluations. They have become a standard in the design and analysis of computer experiments (see e.g. [Sacks et al., 1989], [Jones et al., 1998] and [Santner et al., 2003]). The approach relies on the assumption that the objective function f is a realisation of a Gaussian process Y. In this chapter, we review the state-of-the-art of GP modelling and we discuss the impact of the choice of Y on model construction, prediction results and sequential designs. Section 2.1 is devoted to some generalities on GP modelling and designs of experiments, including targeted designs for a given purpose like sampling f in certain areas of interest. We focus in section 2.2 on non-stationary GP in order to include knowledge on heterogeneous variations of f.

2.1 Framework

2.1.1 Generalities on GP modelling

Basic definitions and key properties

Let us consider a real valued stochastic process $Y = (Y_{\boldsymbol{x}})_{\boldsymbol{x} \in \mathcal{D}}$, i.e. a collection of random variables indexed by $\mathcal{D}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and taking values in the measurable space $(\mathbb{R}, \mathcal{B})$, with $\mathcal{B}$ the Borel σ-algebra of $\mathbb{R}$. At fixed $\omega \in \Omega$, the function $\boldsymbol{x} \to Y_{\boldsymbol{x}}(\omega)$ is called a realisation or a sample path of Y. See for example [Knill, 1994] for an introduction to stochastic processes. A GP is a stochastic process verifying the following.

Definition 1 (Gaussian process)**.** *A stochastic process Y on $\mathcal{D}$ is Gaussian if and only if $\forall n \in \mathbb{N}^\star$, $\forall \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{D}$, and $\forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}$, the finite linear combination of indexed random variables $\sum\limits_{i=1}^{n} \alpha_i Y_{\boldsymbol{x}_i}$ follows a normal distribution.*

From this definition, the mean function $m : \boldsymbol{x} \to \mathbb{E}\left(Y_{\boldsymbol{x}}\right)$, and the covariance function $c : (\boldsymbol{x}, \boldsymbol{x}') \to \operatorname{cov}(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'}) \triangleq \mathbb{E}\left((Y_{\boldsymbol{x}} - m(\boldsymbol{x}))(Y_{\boldsymbol{x}'} - m(\boldsymbol{x}'))\right)$ exist on $\mathcal{D}$ and $\mathcal{D}^2$ respectively. Indeed $Y_{\boldsymbol{x}'}$ and $Y_{\boldsymbol{x}} + Y_{\boldsymbol{x}'}$ have by definition normal distributions and thus finite mean and variance from which $m(\boldsymbol{x})$ and $c(\boldsymbol{x}, \boldsymbol{x}')$ can be derived; in other words, Gaussian processes automatically fulfil the second order stochastic process condition: $\forall \boldsymbol{x} \in \mathcal{D}$, $\mathbb{E}\left(Y_{\boldsymbol{x}}^2\right) < \infty$. Reciprocally, a GP is completely determined in term of finite dimensional distributions by a mean and a covariance function [Rasmussen and Williams, 2006]. To create a GP distribution, one can pick any function from $\mathcal{D}$ to $\mathbb{R}$ as a mean function, but a covariance function c is valid if, and only if, it is a symmetric function (i.e. $\forall \boldsymbol{x}$, $\boldsymbol{x}' \in \mathcal{D}$, $c(\boldsymbol{x}, \boldsymbol{x}') = c(\boldsymbol{x}', \boldsymbol{x})$) and positive definite (i.e. for any choice of $n \in \mathbb{N}^\star$ and weights $\alpha_1, \ldots, \alpha_n$, c verifies $\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j c(\boldsymbol{x}_i, \boldsymbol{x}_j) \geqslant 0$).

In the rest of this section, we give basic properties of a Gaussian process Y with mean $m(\cdot)$ and covariance $c(\cdot, \cdot)$, whose distribution is denoted by $Y \sim \mathcal{GP}(m, c)$.

Link with Gaussian vector conditioning

Consider $Y \sim \mathcal{GP}(m, c)$, for $n \in \mathbb{N}^\star$, a random vector $\boldsymbol{Y}_{1:n} = (Y_{\boldsymbol{x}_1}, \dots, Y_{\boldsymbol{x}_n})^\top$, with $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in \mathcal{D}^n$ is a Gaussian vector[1]. Its mean vector and covariance matrix are:

$$\boldsymbol{m} = (m(\boldsymbol{x}_1), \dots, m(\boldsymbol{x}_n))^\top \text{ and } C = (c(\boldsymbol{x}_i, \boldsymbol{x}_j))_{\substack{i=1,\dots,n \\ j=1,\dots,n}} . \tag{2.1}$$

Let us assume that we observe a realisation $\boldsymbol{y}_{1:n} \in \mathbb{R}^n$ of this vector $\boldsymbol{Y}_{1:n}$ and call this event $\mathcal{A}_n = \{Y_{\boldsymbol{x}_1} = y_1\} \cap \dots \cap \{Y_{\boldsymbol{x}_n} = y_n\}$. For $\boldsymbol{x} \in \mathcal{D}$, the random variable $Y_{\boldsymbol{x}}$ is in general dependent on $\boldsymbol{Y}_{1:n}$ (equivalently 'correlated', in Gaussian case). In fact, knowing the evaluation event $\mathcal{A}_n$ reduces the uncertainty on Y and impacts its distribution. The updated distribution of $Y_{\boldsymbol{x}}$ knowing $\mathcal{A}_n$ is called conditional distribution of $Y_{\boldsymbol{x}}$ given $\mathcal{A}_n$ and can be analytically computed using the following property of Gaussian vectors.

Proposition 1 (Gaussian vector conditioning)**.** *Let* $\boldsymbol{Y} = \left(\boldsymbol{Y}^{(a)}, \boldsymbol{Y}^{(b)}\right)^\top$ *be a Gaussian vector with* $\mathbb{E}\left(\boldsymbol{Y}^{(a)}\right) = \boldsymbol{m}^{(a)}$, $\mathbb{E}\left(\boldsymbol{Y}^{(b)}\right) = \boldsymbol{m}^{(b)}$, $\operatorname{cov}\left(\boldsymbol{Y}^{(a)}\right) = C_{a,a}$ *invertible,* $\operatorname{cov}\left(\boldsymbol{Y}^{(b)}\right) = C_{b,b}$ *and* $\operatorname{cov}\left(\boldsymbol{Y}^{(a)}, \boldsymbol{Y}^{(b)}\right) = C_{a,b}$*. Then the conditional distribution of* $\boldsymbol{Y}^{(b)}$ *knowing the event* $\boldsymbol{Y}^{(a)} = \boldsymbol{y}^{(a)}$*, is also a multivariate Gaussian distribution, with mean and covariance matrix*

$$\boldsymbol{m}_{b|a}\left(\boldsymbol{Y}^{(a)}\right) := \mathbb{E}\left(\boldsymbol{Y}^{(b)} \middle| \boldsymbol{Y}^{(a)}\right) = \boldsymbol{m}^{(b)} + C_{a,b}^\top C_{a,a}^{-1}\left(\boldsymbol{Y}^{(a)} - \boldsymbol{m}^{(a)}\right) \tag{2.2}$$

and

$$C_{b|a} := \operatorname{cov}\left(\boldsymbol{Y}^{(b)} \middle| \boldsymbol{Y}^{(a)}\right) = C_{b,b} - C_{a,b}^\top C_{a,a}^{-1} C_{a,b}. \tag{2.3}$$

When the mean and the covariance are known, these formulas can be used to calculate directly the mean function m_n and the covariance function c_n of Y conditioned on n observations. We have for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$

$$m_n(\boldsymbol{x}) = \mathbb{E}\left(Y_{\boldsymbol{x}} \middle| \mathcal{A}_n\right) = m(\boldsymbol{x}) + \boldsymbol{c}_n(\boldsymbol{x})^\top C^{-1}\left(\boldsymbol{y}_{1:n} - \boldsymbol{m}\right), \tag{2.4}$$

$$c_n(\boldsymbol{x}, \boldsymbol{x}') = \operatorname{cov}\left(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'} \middle| \mathcal{A}_n\right) = c(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{c}_n(\boldsymbol{x})^\top C^{-1} \boldsymbol{c}_n(\boldsymbol{x}') \tag{2.5}$$

with $\boldsymbol{c}_n : \boldsymbol{x} \to (c(\boldsymbol{x}, \boldsymbol{x}_i))_{i=1,\dots,n}^\top$ and $\boldsymbol{m}$, C as in eq. (2.1).

[1] A Gaussian vector is a random vector such that any linear combination of its components is normally distributed.

Basics on building GP models

Example of a GP conditioned on data points. Using GP conditioning, we can already illustrate our first example of GP model. Let us assume that f is a realisation of Y of given mean and covariance functions m and c. With this hypothesis, evaluated values $\boldsymbol{y}_{1:n}$ at some points $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ are considered as a realisation of the vector $\boldsymbol{Y}_{1:n}$. Assuming for simplicity that m and c are given, they often involve parameters that are actually estimated on $\mathcal{A}_n$ in practice. The GP conditioned on an evaluation set $\mathcal{A}_n$ has distribution $\mathcal{GP}(m_n, c_n)$ as in eq. (2.4) and it can be used as a probabilistic model of f. The conditional mean m_n is used as a prediction and c_n provides a measure of the prediction uncertainty, in particular with the prediction standard deviation $\sigma_n : \boldsymbol{x} \to \sqrt{c_n(\boldsymbol{x}, \boldsymbol{x})}$.

Figure 2.1 shows mean, variance and some sample paths of a GP for a one-dimensional test case. We consider $\mathcal{D} = [0, 1]$, m the null function on $\mathcal{D}$ and c a stationary Matérn covariance with smoothness parameters $\nu = 5/2$ [Stein, 1999, Rasmussen and Williams, 2006]

$$k^{\text{`}5/2\text{'}}_{\theta,\sigma} : h \to \sigma^2 \left(1 + \sqrt{5}\frac{h}{\theta} + \frac{5}{3}\left(\frac{h}{\theta}\right)^2\right) \exp\left(-\sqrt{5}\frac{h}{\theta}\right), \tag{2.6}$$

fixed correlation length $\theta = 1/10$, and standard deviation $\sigma = 1$ (see later in this section 2.1.1 for a discussion on covariance parametrisation). The GP is then conditioned on four evaluations points arbitrarily taken. We observe a reduction of the variance around the evaluations and an interpolation of the evaluation by the mean and by the sample paths from the conditional distribution. For an illustrations of a bivariate model, see e.g. fig. 2.4. We see now some approach for accounting for the uncertainty on m and c.

From prior assumptions to predictions. The term *prior* normally refers to probability distributions assumed before taking the evaluations into account, as opposed to *posterior* distributions. The model and its quality strongly depend on the determination of the prior Gaussian process Y, especially if few evaluations are available. Often in practice, the distribution of Y is parametrised by a (real-valued) vector of unknown parameters $\boldsymbol{\theta} \in \Theta$, where Θ is a multidimensional parametric space. The vector $\boldsymbol{\theta}$ is then estimated with the data set. In a pure Bayesian model, $\boldsymbol{\theta}$ follows a given probability distribution which impacts the posterior distribution of Y via the Bayes formula [Rasmussen and Williams, 2006]. In contrast, we adopt in this thesis the empirical Bayes view-point as in [Sacks et al., 1989, Roustant et al., 2012]. In this

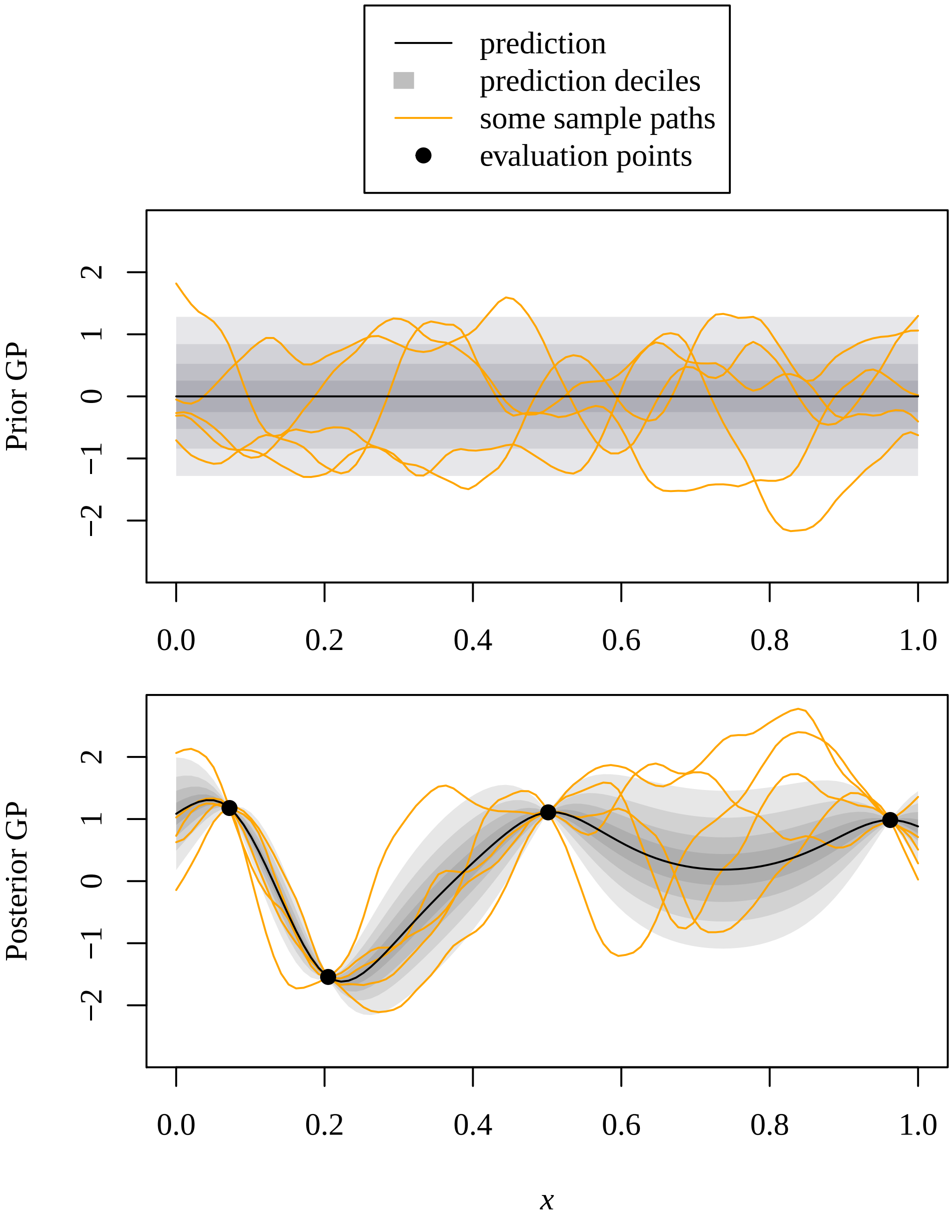


Figure 2.1: Gaussian process conditioning.

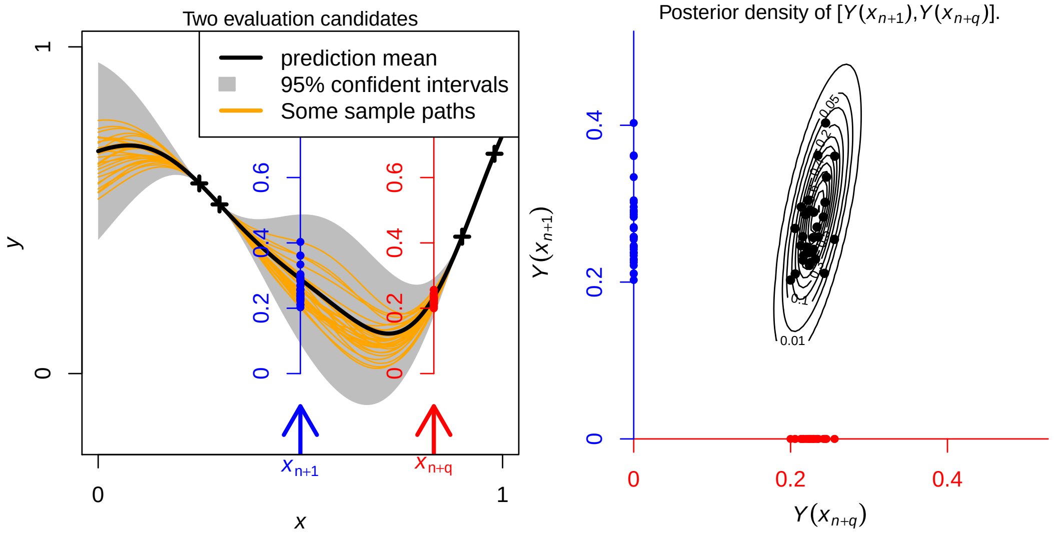


Figure 2.2: GP model (right) and posterior join distribution at two candidate evaluation points.

setting, one must first get the formulas of the posterior distribution calculated with a deterministic arbitrary $\boldsymbol{\theta}$. Then the GP model is obtained by plugging in these formulas a value of $\boldsymbol{\theta}$ estimated from the data, e.g. defined by cross-validation minimisation or by maximum likelihood (see [Park and Baek, 2001] and paragraph 'Parameter estimation').

Analytical formula for a Gaussian posterior distribution of Y are tractable if we consider the following formula determining the prior

$$Y_{\boldsymbol{x}} = \boldsymbol{\beta}^{\top}\boldsymbol{g}(\boldsymbol{x}) + Z_{\boldsymbol{x}} \tag{2.7}$$

where $\boldsymbol{g}$ is a function from $\mathcal{D}$ to $\mathbb{R}^p$, $\boldsymbol{\beta} = \boldsymbol{\theta}$ a random vector of unknown trend parameters following an improper prior distribution[2], and is Z a given centred GP. For this setting, corresponding to the case of 'universal kriging' [Matheron, 1973, Handcock and Stein, 1993], formulas for the posterior distribution are [Roustant et al., 2012]

$$\mathbb{E}\left(Y_{\boldsymbol{x}} \middle| \mathcal{A}_n\right) = \boldsymbol{g}(\boldsymbol{x})^{\top}\widehat{\boldsymbol{\beta}} + \boldsymbol{c}(\boldsymbol{x})^{\top}C^{-1}(\boldsymbol{y}_{1:n} - G\widehat{\boldsymbol{\beta}}) \tag{2.8}$$

$$\begin{aligned}\operatorname{cov}\left(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'} \middle| \mathcal{A}_n\right) = {} & c(\boldsymbol{x},\boldsymbol{x}') - \boldsymbol{c}(\boldsymbol{x})^{\top}C^{-1}\boldsymbol{c}(\boldsymbol{x}') \qquad (2.9)\\ & +(\boldsymbol{g}(\boldsymbol{x})^{\top} - \boldsymbol{c}(\boldsymbol{x})^{\top}C^{-1}G)^{\top}(G^{\top}C^{-1}G)^{-1}(\boldsymbol{g}(\boldsymbol{x}')^{\top} - \boldsymbol{c}(\boldsymbol{x}')^{\top}C^{-1}G)\end{aligned}$$

[2] A improper prior can be seen as a limit of uniform distributions when their support converges to $\mathbb{R}^p$. This type of prior distribution are used under conditions that insure a proper posterior distribution (see [Helbert et al., 2009]).

with $\widehat{\boldsymbol{\beta}} = \left(G^\top C G\right)^{-1} G^\top C^{-1} \boldsymbol{y}_{1:n}$, $\boldsymbol{c}(\boldsymbol{x})$ is defined as in *eq.* (2.4) (i.e. $\boldsymbol{c} : \boldsymbol{x} \to (c(\boldsymbol{x}, \boldsymbol{x}_i))_{i=1,\dots,n}^\top$) and G is the $n \times p$ matrix $\left(\boldsymbol{g}(\boldsymbol{x}_1), \dots, \boldsymbol{g}(\boldsymbol{x}_n)\right)^\top$. In a Bayesian interpretation, these formulas derive from the assumption that the process Z has a given distribution $\mathcal{GP}(0_{\mathbb{R}^\mathcal{D}}, c)$ (with $0_{\mathbb{R}^\mathcal{D}}$ the null function on $\mathcal{D}$). In practice $\boldsymbol{\theta}$ contains not only the trend coefficients $\boldsymbol{\beta}$, but also covariance parameters of Z and possible other model parameters. In contrast to $\boldsymbol{\beta}$, the estimations of other parameters rarely use closed form formulas and require numerical estimation approaches (see later e.g. for cross-validation and maximum likelihood approaches in an empirical Bayes manner).

The subcase named 'ordinary kriging' corresponds to p equals 1 and g_1 is the constant function equal to 1. The matrix G is then a vector of size n with all components equal 1, denoted $\mathbf{1}_n$, and $\widehat{\boldsymbol{\beta}}$ is a real value $\widehat{\beta} = \frac{1}{\mathbf{1}_n^\top C^{-1} \mathbf{1}_n} \mathbf{1}_n^\top C^{-1} \boldsymbol{y}_{1:n}$.

We conclude this part with an illustration of an example of ordinary kriging (fig. 2.2). The settings are exactly the same as for fig. 2.1 except two differences: the mean value is not fixed to zero but to an unknown constant (using an improper prior as in Helbert et al. [2009]), and the covariance parameters are estimated by maximum likelihood following the procedure explained in the following paragraph. This figure also display the distribution of a vector $(Y_{\boldsymbol{x}_{n+1}}, Y_{\boldsymbol{x}_{n+2}})$ for a batch of $q = 2$ arbitrary points $\boldsymbol{x}_{n+1}$, $\boldsymbol{x}_{n+q} \in \mathcal{D}$.

Cross-validation. We now review methods for estimating a parametrised model from a data set. The estimation by cross-validation aims at minimising prediction errors, for instance the integrated squared error Δ

$$\Delta(\boldsymbol{\theta}) = \int_{\mathcal{D}} \left(f(\boldsymbol{u}) - m_{n,\boldsymbol{\theta}}(\boldsymbol{u})\right)^2 \mathrm{d}\boldsymbol{u}, \tag{2.10}$$

with $m_{n,\boldsymbol{\theta}}$ the model prediction built from n observations. Its dependence on the parameter vector $\boldsymbol{\theta}$ is emphasised via the subscript. As f is only partially known, Δ is approximated using errors at each evaluation point $\boldsymbol{x}_i$ when this evaluation is withdrawn from the training data set. A general *k-fold* cross-validation requires first to partition the evaluations into k groups (or folds). We denote with $\mathcal{I} : \{1, \dots, n\} \to \{1, \dots, k\}$ the function giving the group index of each evaluation. The model estimator is then

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(m_{n,\boldsymbol{\theta}}^{(i)}(\boldsymbol{x}_i) - y_i \right)^2 \right) \tag{2.11}$$

with $m_{n,\boldsymbol{\theta}}^{(i)}(\boldsymbol{x}_i)$ the prediction at point i without the evaluations of the group $\mathcal{I}(i)$, i.e. without evaluations with index in the preimage $\mathcal{I}^{-1}(\mathcal{I}(i))$.

When $k = n$ (corresponding to $\mathcal{I}(i) = i$, $i = 1, \dots, n$), the prediction at point i is made after removing only the i^{th} evaluation. This setting is called *leave-one-out* cross validation. The constant k, which has an impact on $\widehat{\boldsymbol{\theta}}$, remains a parameter that needs to be tuned. As the number of folds increase, the computational cost of the cross-validation process increase linearly with the number of folds. Putting these constraints aside, the choice of k involves for $\boldsymbol{\theta}$ a bias-variance trade-off discussed e.g. in [James et al., 2013]. For a more general overview on cross-validation, see Arlot [2008] and references therein.

Maximum likelihood. Maximum likelihood estimation (MLE) consists of finding a parameter vector $\boldsymbol{\theta}$ maximising the (logarithm of the) likelihood function given the evaluations. For given parameter values $\boldsymbol{\theta}$, the likelihood is defined here as the prior probability density of the Gaussian vector $\boldsymbol{Y}_{1:n}$ evaluated at the observed values $\boldsymbol{y}_{1:n}$:

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}_{1:n}) = \varphi_{n,C(\boldsymbol{\theta})}(\boldsymbol{y}_{1:n} - \boldsymbol{m}(\boldsymbol{\theta})) \tag{2.12}$$

with $\boldsymbol{m}(\boldsymbol{\theta})$ and $C(\boldsymbol{\theta})$ (as in eq. (2.1)) depending on $\boldsymbol{\theta}$ and with $\varphi_{n,\Sigma}$ the probability density function of $\mathcal{N}(0, \Sigma)$, a centred n-variate normal distribution with covariance matrix Σ. In this work, parameter estimation by maximum likelihood is mainly performed via the R package *kergp* [Deville et al., 2015]; the numerical optimisation relies on the BFGS algorithm [Battiti and Masulli, 1990] with one or several initial evaluations.

Classical covariance structures

In the absence of specific prior assumption on f, it is common to take stationary GP distributions [Stein, 1999] as prior. The finite-dimensional distributions of stationary GP are invariant by translation: a GP is stationary if and only if, for any $\boldsymbol{h}$, $\boldsymbol{x} \in \mathbb{R}^d$, with $(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) \in \mathcal{D} \times \mathcal{D}$, the distribution of outputs $(Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}+\boldsymbol{h}})^\top$, does not depend on $\boldsymbol{x}$, but only on $\boldsymbol{h}$. A covariance function c is said to be stationary when it defines a stationary centred[3] GP distribution, meaning that there exists a real-valued function g on $\mathbb{R}^d$ such that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$:

$$c(\boldsymbol{x}, \boldsymbol{x}') = g(\boldsymbol{x} - \boldsymbol{x}'). \tag{2.13}$$

When a covariance depends only on the Euclidean norm $||\boldsymbol{x} - \boldsymbol{x}'||$, i.e. written $c(\boldsymbol{x}, \boldsymbol{x}') = k(||\boldsymbol{x} - \boldsymbol{x}'||)$, it is called ‘isotropic’. Not every function k guarantees

[3] i.e. the mean function is constant equal to zero.

that c is a covariance function: the function k is named 'positive definite radial basis function on $\mathcal{D}$' if c is positive definite on $\mathcal{D}$. The function from the Matérn family with parameter $\nu = 5/2$ (see eq. (2.6)) and the Gaussian (or square exponential) function $k^{`\infty'} : h \to \exp(-h^2)$ are common positive definite radial basis functions. The Matérn class is quite popular not only for radial Matérn kernels [Rasmussen and Williams, 2006], but also for its tensor product counterparts [Roustant et al., 2012], where c is formulated on $\mathcal{D}^2$ as

$$c(\boldsymbol{x}, \boldsymbol{x}') = k^{`5/2'}_{\theta_1,\sigma_1}(|x_1 - x'_1|) \times \ldots \times k^{`5/2'}_{\theta_d,\sigma_d}(|x_d - x'_d|) \tag{2.14}$$

with $\theta_1, \ldots, \theta > 0$ and $\sigma_1, \ldots, \sigma_d \geqslant 0$. A reason for the success of Matérn kernels is its tunable smoothness, with the parameter ν controlling the order of (almost sure) differentiability of the associated GP realisations (see details on differentiability later in this section).

Given an isotropic kernel on $\mathbb{R}^d$, geometric anisotropy can be easily generated by replacing the Euclidean distance with a distance sometimes called Mahalanobis distance

$$\mathrm{dist}_\Gamma(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{(\boldsymbol{x} - \boldsymbol{x}')^\top \Gamma (\boldsymbol{x} - \boldsymbol{x}')}, \tag{2.15}$$

where Γ is a symmetric definite matrix. With expensive evaluations, it is important to keep the number of model parameters moderate. But in general anisotropic GP models, the dimension of $\boldsymbol{\theta}$ increases quadratically with d. In general, geometric anisotropy requires to parametrise a rotation ($d(d-1)/2$ parameters) and length-scale parameters for each axis ($d(d-1)/2$ parameters in total). With low rank Γ, it is possible to reduce parameters to a linear number with dimensionality reduction (see e.g [Rasmussen and Williams, 2006]). One can also consider parsimonious multidimensional non-linear regression like the Single Index Model (SIM) [Brillinger, 1977]. In the framework of Gaussian Process models (see GP-SIM, [Choi et al., 2011, Gramacy and Lian, 2012b]), the prior covariance is defined from an univariate covariance $k_{\boldsymbol{b}}$, parametrised by a vector $\boldsymbol{b}$, chained with a scalar product with a vector $\boldsymbol{a} \in \mathbb{R}^d$:

$$c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\boldsymbol{b}}(\boldsymbol{a}^\top \boldsymbol{x}, \boldsymbol{a}^\top \boldsymbol{x}'). \tag{2.16}$$

If $k_{\boldsymbol{b}}$ is stationary, $c_{\boldsymbol{\theta}}$ is also stationary. With this covariance function, the dimension of $\boldsymbol{\theta} = \{\boldsymbol{b}, \boldsymbol{a}\}$ increases affinely in d with slope 1. With a fixed $\boldsymbol{\theta}$, this model has a constant prior covariance on any subspace of the type $H_1 \times H_2$, with H_1, H_2 hyperplanes with normal vector $\boldsymbol{a}$. Relaxing this constraint, the multiple index model is an extension proposed by [Xia, 2008]. It uses a more complex r-variate covariance function for $k_{\boldsymbol{b}}$, $r \in \mathbb{N}^\star$, and extends the scalar product to a matrix product:

$$c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\boldsymbol{b}}\left(A\boldsymbol{x}, A\boldsymbol{x}'\right) \tag{2.17}$$

where A is a $r \times d$ matrix and $k_{\boldsymbol{b}}$ a parametrised r dimensional positive definite kernel. If $k_{\boldsymbol{b}}$ is isotropic on $\mathbb{R}^r$, say with radial basis function $k^{\text{‘5/2’}}_{b_1,b_2}$, b_1, $b_2 \geqslant 0$, it follows that $c_{\boldsymbol{\theta}}$ is anisotropic on $\mathbb{R}^d$ with Mahalanobis distance matrix $\Gamma = A^\top A$:

$$c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k^{\text{‘5/2’}}_{b_1,b_2}\left(\sqrt{(\boldsymbol{x}-\boldsymbol{x}')A^\top A(\boldsymbol{x}-\boldsymbol{x}')}\right). \tag{2.18}$$

Differentiability of GPs

The smoothness of GPs are linked in particular to the properties of their covariance functions. Let us now briefly review some definitions and properties about GP differentiability.

Mean squared regularity. For second order stochastic processes, and in particular GPs, mean squared continuity and differentiability is defined as follow.

Definition 2. *Given a point $\boldsymbol{x} \in \mathcal{D}$, a GP Z is said to be mean-square continuous at $\boldsymbol{x}$ if $\mathbb{E}\left(Z_{\boldsymbol{x}}^2\right) < +\infty$ and*

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \left[\mathbb{E}\left((Z_{\boldsymbol{x}+\boldsymbol{h}} - Z_{\boldsymbol{x}})^2\right)\right] = 0. \tag{2.19}$$

Furthermore, mean squared differentiability of Z at a point $\boldsymbol{x} \in \mathcal{D}$ in the i^{th} canonical direction is established by the existence of a random variable $U_{i,\boldsymbol{x}}$ of order 2 ($\in L^2(\Omega, \mathbb{P})$) such that

$$\lim_{h\to 0}\left[\mathbb{E}\left(\left(\frac{Z_{\boldsymbol{x}+he_i} - Z_{\boldsymbol{x}}}{h} - U_{i,\boldsymbol{x}}\right)^2\right)\right] = 0. \tag{2.20}$$

We write $U_{i,\boldsymbol{x}} = \frac{\partial}{\partial t_i} Y_{\boldsymbol{t}}\Big|_{\boldsymbol{t}=\boldsymbol{x}}$ or simply $\frac{\partial}{\partial x_i} Y_{\boldsymbol{x}}$, and $(U_1, \ldots, U_d)^\top = \nabla_{\boldsymbol{t}}\left[Y_{\boldsymbol{t}}\right](\boldsymbol{x})$, or simply $\nabla Y_{\boldsymbol{x}}$.

Continuity and differentiability of a GP is closely related to a similar regularity for the covariance function as stated in the following proposition.

Proposition 2 (Characterisation of mean square regularity)**.** *Let Z be a GP on $\mathcal{D}$ with continuous mean function m and covariance function c. Z is mean-square continuous if and only if*

$$\boldsymbol{x} \to c(\boldsymbol{x}, \boldsymbol{x}) \tag{2.21}$$

is continuous on $\mathcal{D}$.

Assuming the differentiability of m and the existence for all $i = 1, \ldots, d$, $\boldsymbol{x} \in \mathcal{D}$ of derivatives $\frac{\partial^2}{\partial t_i \partial t_i'} c(\boldsymbol{t}, \boldsymbol{t}')\Big|_{\boldsymbol{t}=\boldsymbol{t}'=\boldsymbol{x}}$, then for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$, $k, \ell \in 1, \ldots, d$, we have

$$\mathbb{E}\left(\nabla Y_{\boldsymbol{x}}\ \right) = \nabla m(\boldsymbol{x}), \ \ \textit{and} \tag{2.22}$$

$$\operatorname{cov}\left(\frac{\partial Y_{\boldsymbol{x}}}{\partial x_k}, \frac{\partial Y_{\boldsymbol{x}'}}{\partial x_\ell}\right) = \left.\frac{\partial^2 c(\boldsymbol{t}, \boldsymbol{t}')}{\partial t_k \partial t_\ell'}\right|_{\boldsymbol{t}=\boldsymbol{x}, \boldsymbol{t}'=\boldsymbol{x}'}. \tag{2.23}$$

A Gaussian process with such property induces a distribution for their derivatives. The joint distribution between Y and its derivatives can be derived, see e.g. Wu et al. [2017]. In particular, for every $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{D}$, integer $\ell \leqslant n$, and indices $J = (j_1, \ldots, j_\ell) \in \{1, \ldots, d\}^\ell$, the Gaussian vector $\left(\frac{\partial Y_{\boldsymbol{x}_1}}{\partial x_{j_1}}, \ldots, \frac{\partial Y_{\boldsymbol{x}_p}}{\partial x_{j_\ell}}, Y_{\boldsymbol{x}_{\ell+1}}, \ldots, Y_{\boldsymbol{x}_n}\right)^\top$ is distributed as

$$\mathcal{N}\left(\left(\frac{\partial m(\boldsymbol{x}_1)}{\partial x_{j_1}}, \ldots, \frac{\partial m(\boldsymbol{x}_\ell)}{\partial x_{j_\ell}}, m(\boldsymbol{x}_{\ell+1}), \ldots, m(\boldsymbol{x}_n)\right)^\top, \begin{pmatrix} {}^{\nabla^2}C & {}^{\nabla}C \\ {}^{\nabla'}C & C \end{pmatrix}\right) \tag{2.24}$$

$$\begin{aligned}
\text{with} \qquad C &= \left(c(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)_{\substack{i=\ell+1,\ldots,n \\ i'=\ell+1,\ldots,n}} \\
{}^{\nabla}C &= \left(\left.\frac{\partial c(\boldsymbol{t}, \boldsymbol{x}_i)}{\partial t_{j_i}}\right|_{\boldsymbol{t}=\boldsymbol{x}_i}\right)_{\substack{i=1,\ldots,\ell \\ i'=\ell+1,\ldots,n}} \\
{}^{\nabla'}C &= \left(\left.\frac{\partial c(\boldsymbol{x}_i, \boldsymbol{t}')}{\partial t_{j_{i'}}}\right|_{\boldsymbol{t}'=\boldsymbol{x}_{i'}}\right)_{\substack{i=\ell+1,\ldots,n \\ i'=1,\ldots,\ell}} \\
{}^{\nabla^2}C &= \left(\left.\frac{\partial^2 c(\boldsymbol{t}, \boldsymbol{t}')}{\partial t_{j_i} \partial t_{j_{i'}}'}\right|_{\substack{\boldsymbol{t}=\boldsymbol{x}_i \\ \boldsymbol{t}'=\boldsymbol{x}_{i'}}}\right)_{\substack{i=1,\ldots,\ell \\ i'=1,\ldots,\ell}}.
\end{aligned}$$

This allows for Bayesian conditioning given evaluations of a function or its derivatives.

Sample path differentiability. Although mean square properties are easier to derive from the regularity of the covariance, the interpretations of sample path properties are more straightforward as they directly inform on the assumptions on f.

Definition 3. *Given a point $\boldsymbol{x} \in \mathcal{D}$, a Gaussian process Y is said to be sample path continuous at $\boldsymbol{x}$ if*

$$\mathbb{P}\left(\lim_{\boldsymbol{h}\to\boldsymbol{0}} Z_{\boldsymbol{x}+\boldsymbol{h}} = Z_{\boldsymbol{x}}\right) = 1. \tag{2.25}$$

Furthermore, sample path differentiability of Y at a point $\boldsymbol{x} \in \mathcal{D}$ in the i^{th} canonical direction is established by the existence of a random variable $U_{i,\boldsymbol{x}}$ such that

$$\mathbb{P}\left(\lim_{h\to 0} \frac{Z_{\boldsymbol{x}+h\boldsymbol{e}_i} - Z_{\boldsymbol{x}}}{h} = U_{i,\boldsymbol{x}}\right) = 1. \tag{2.26}$$

As the mean square partial derivatives are equal to the corresponding sample path derivatives with probability one [Doob, 1953] we also write $U_{i,\boldsymbol{x}} = \frac{\partial}{\partial x_i} Y_{\boldsymbol{x}}$, and $\nabla Y_{\boldsymbol{x}} = (U_1, \dots, U_d)^\top$.

Proposition 3 (Sufficient conditions for sample path differentiability)**.** *Let Y be a separable[4], on an open subset $\mathcal{D}$ of $\mathbb{R}^d$, with covariance function c. If for all $i = 1, \dots, d$, $\frac{\partial c}{\partial x_i \partial x_i'}$ exists on $\mathcal{D}^2$, and if for some $b, h, \varepsilon > 0$, it holds that for all $i = 1, \dots, d$, and for all $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathcal{D}$ with $||\boldsymbol{x} - \boldsymbol{x}'|| \leqslant \varepsilon$,*

$$\left.\frac{\partial^2 c(\boldsymbol{t}, \boldsymbol{t}')}{\partial t_i \partial t_{i'}'}\right|_{\substack{\boldsymbol{t}=\boldsymbol{x}\\ \boldsymbol{t}'=\boldsymbol{x}}} + \left.\frac{\partial^2 c(\boldsymbol{t}, \boldsymbol{t}')}{\partial t_i \partial t_{i'}'}\right|_{\substack{\boldsymbol{t}=\boldsymbol{x}'\\ \boldsymbol{t}'=\boldsymbol{x}'}} - 2\left.\frac{\partial^2 c(\boldsymbol{t}, \boldsymbol{t}')}{\partial t_i \partial t_{i'}'}\right|_{\substack{\boldsymbol{t}=\boldsymbol{x}\\ \boldsymbol{t}'=\boldsymbol{x}'}} \leqslant \frac{b}{\left|\ln ||\boldsymbol{x} - \boldsymbol{x}'||\right|^{1+h}} \tag{2.27}$$

then the sample paths of Y are $\mathcal{C}^1(\mathcal{D})$ with probability one.

See [Scheuerer [2009], p. 55] for a proof. This property says that a GP is sample path differentiable if its mean square partial derivatives exist (through the condition on c) and are sample path continuous (through the inequality condition (2.27), see Adler [2010]).

[4] In general the distribution of a stochastic process does not determine the sample paths properties, as discussed in [Doob, 1953] centred Gaussian process, and in e.g. Scheuerer [2009]. We assume in this thesis that the Gaussian processes are separable, meaning that the distribution determines the properties of sample paths (see [Scheuerer [2009], section 5.2] for a definition).

Wavelet analysis and its application to stochastic processes

The continuous wavelet transform [Daubechies, 1992] of $y \in L^2(\mathbb{R})$ with respect to an admissible[5] wavelet $\psi \in L^2(\mathbb{R})$ consists of scalar products between y with translated and dilated instances of ψ:

$$\forall(\tau, s) \in \mathbb{R}^2,\ \mathcal{W}_y(\tau, s) = \int_{\mathbb{R}} y(x)\psi_{\tau,s}(x)\mathrm{d}x, \tag{2.28}$$

where $\psi_{\tau,s} : x \to \frac{1}{\sqrt{h^s}}\psi\left(\frac{x-\tau}{h^s}\right)$ with $h \in]0, \infty[$. For each values (τ, s), $\mathcal{W}_y(\tau, s)$ is the wavelet coefficient of y related to the scale h^{-s} and the position τ.

Concerning wavelets methods throughout this thesis, most of the time $d = 1$ is assumed, except for the presentation of some general concepts or for the bivariate case study (section 5.3.2). The wavelet transform is computed using an analytic derivative of the Gaussian wavelet, as in [Omer and Torresani, 2016], defined by its Fourier transform $\hat{\psi}$ which vanishes for negative frequencies

$$\hat{\psi}(u) = ue^{-u^2} \text{ for } u \geqslant 0, \hat{\psi}(u) = 0 \text{ for } u < 0. \tag{2.29}$$

Figure 2.3 illustrates a wavelet transform of a simple function with varying local frequency.

The wavelet transform can be applied to the sample paths of a stochastic process $Y \sim \mathcal{GP}(m, c)$. Guérin [2000] gives sufficient conditions to ensure the almost sure existence of the resulting transform (here $d = 1$):

- Y is a second order mean square continuous stochastic process,
- there exists $r > 0$ such that $\mathbb{E}(Y_x Y_{x'}) = O(|xx'|^r)$ at infinity.
- the wavelet functions belong to the Schwartz space, i.e. the space of C^∞ functions vanishing rapidly at infinity:

$$\mathcal{S}(\mathbb{R}) : \left\{\psi \in \mathcal{C}^\infty(\mathbb{R})|\, \forall\alpha \geqslant 0, \beta \in \mathbb{N},\ \sup_{t\in\mathbb{R}}\left(|t^\alpha \psi^{(\beta)}(t)|\right) < \infty\right\}.$$

From there it follows that for all $\tau,\ s \in \mathbb{R}$,

$$\int_{\mathbb{R}} \mathbb{E}\,|\psi_{\tau,s}(x)Y_x|\,\mathrm{d}x < \infty, \tag{2.30}$$

which ensures that the sample paths of the process $\left(\psi_{\tau,s}(x)Y_x\right)_{x\in\mathcal{D}}$ are integrable, leading to the definition for all $\tau, s \in \mathbb{R}^2$ of the random variable $\mathcal{W}_Y(\tau, s)$.

[5] i.e. $\int_{\mathbb{R}} \psi(t)\mathrm{d}t = 0$

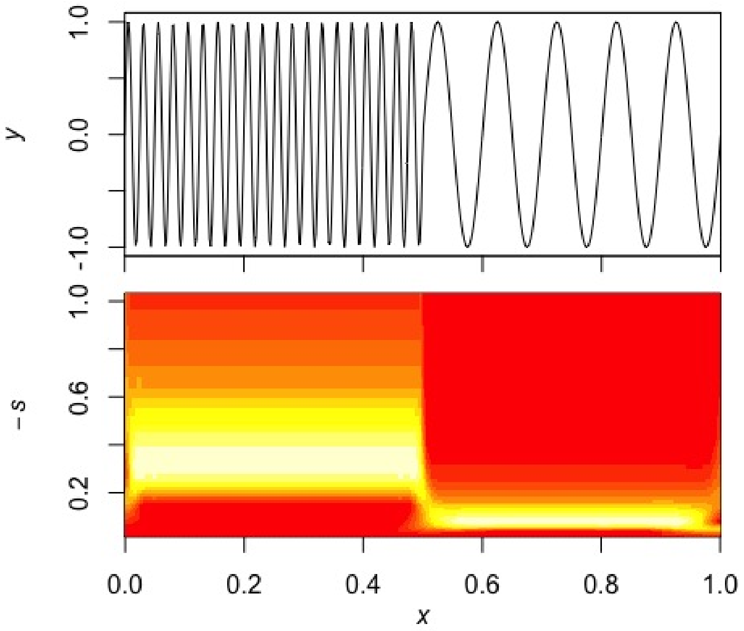


Figure 2.3: Wavelet transform of $x \to \sin(80\pi x)\mathbb{1}_{[0,0.5]}(x)+\sin(20\pi x)\mathbb{1}_{]0.5,1]}(x)$, with $h = 1/2$.

We essentially focus on cases where Y is a Gaussian Process with almost surely continuous paths, so that $\mathcal{W}_Y(\tau, s)$ exists for all τ, s and $(\mathcal{W}_Y(\tau, s))_{(\tau,s)\in\mathbb{R}^2}$ is itself a Gaussian process. The wavelet analysis of a Gaussian process is further exploited for the estimation of the warping involved in non-stationary GP models in section 3.4.

2.1.2 Generalities on design of computer experiments

A design of experiments is a set of points in the input space $\mathcal{D}$ locating the experiment evaluations. By extension, 'design of experiments' refers to the decision rules that determines the evaluation locations. In a case of expensive evaluations, the choice of design is crucial.

Model-free designs

When there is almost no information available on the studied function f, 'space-filling' designs spread out as much as possible a given number of evaluations across the input space. Such designs cover the whole input space in an equal (or uniform) manner. Model-free designs may be used as a first sampling for building an initial model. The rest of the evaluations is then allocated according to a model-based design of experiments (see in next section). In this thesis, such designs, filling the input space, are obtained using the R package *DiceDesign*, and are generated with the algorithm of 'Latin hypercube sampling, optimised with a maximin distance criterion' [Dupuy et al., 2015]. See [Pronzato and Müller, 2011] for a more detailed review on the space-filling designs and their comparisons.

GP-based sequential designs

Let us consider a GP model built on a initial design of size n_0. This first approximation of f is a starting point of many sequential sampling methods. In these methods, the sequential design itself is a loop incrementing n, the current number of evaluations $n = n_0 + 1, \dots, N$. If q experiments can be performed at the same time, we can consider a (synchronised[6]) parallel design of experiment in which n is incremented by q evaluations at each step.

Sequential sampling is typically driven by the optimisation of a family of infill criteria $J_{n,q}$ coupled with updating model parameters at each iteration. More precisely, the next batch of evaluation points $X = \{\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_{n+q}\}$ are selected as

$$X \in \operatorname*{argmax}_{\breve{X} \in \mathcal{D}^q} J_{n,q}\left(\breve{X}\right). \tag{2.31}$$

$J_{n,q}$ depends on the n subsequent evaluations: it is defined in terms of the mean m_n and the covariance c_n of the posterior GP with n evaluations.

Variance-based designs

A first idea for a criterion is to evaluate the point from which the prediction has the highest posterior variance $\boldsymbol{x} \to c_n(\boldsymbol{x}, \boldsymbol{x})$, 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, and allocate evaluations to unexplored regions. First MSE, defined for $q = 1$, chooses the next evaluation in points where the variance on the prediction is high:

$$J_{n,1}^{\text{MSE}}(\boldsymbol{x}) = c_n(\boldsymbol{x}, \boldsymbol{x}). \tag{2.32}$$

[6] meaning that the evaluations of f are treated as if they require the same time.

In the case of deterministic evaluations, the MSE criterion can be reformulated in maximin terms by changing the metric on $\mathcal{D}$ to the canonical metric of the GP model covariance c_n, i.e. considering the distance $d(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{c_n(\boldsymbol{x}, \boldsymbol{x}) + c_n(\boldsymbol{x}', \boldsymbol{x}') - 2c_n(\boldsymbol{x}, \boldsymbol{x}')}$. Thus MSE maximisation amounts to maximising a minimal distance to available design points, taking a distance that accounts for covariances given by the model rather than the Euclidean distance.

As a generalisation for the multipoint case with $q > 1$, $X = \{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q}\}$, we can consider the matrix of the posterior covariances $(c_n(\,\boldsymbol{x}_i, \boldsymbol{x}_j\,))_{i,j=1,\ldots,q}$ instead of the variance $c_n(\boldsymbol{x}, \boldsymbol{x})$. This matrix is then chained with a real-valued function to define a criterion. Taking the trace of the matrix would lead to trivial optimum, with $\boldsymbol{x}_1 = \boldsymbol{x}_2 = \ldots = \boldsymbol{x}_q = \boldsymbol{x}^*$, where $\boldsymbol{x}^*$ is a maximiser of the MSE $\boldsymbol{x} \rightarrow c_n(\boldsymbol{x}, \boldsymbol{x})$. In contrast, we want that our criterion makes a trade-off between first evaluating in the regions with highest model uncertainty for each $\boldsymbol{x}_i$, and second spreading out these points in $\mathcal{D}$. To get such compromise between maximising the variance at each point and maximising their distance from each other, the determinant is proposed. The criterion $J_{n,q}^{\mathrm{MSE}}$ is then

$$J_{n,q}^{\mathrm{MSE}}(X) = \det\left((c_n(\,\boldsymbol{x}_i, \boldsymbol{x}_j\,))_{i,j=1,\ldots,q}\right)$$

The usage of the determinant leads to maximising the hypervolume of an ellipsoid of isoprobability of the Gaussian vector $(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_q))^\top$. This criteria is equivalent to the entropy maximisation of Shewry and Wynn [1987] (see [Fang et al., 2006, Pronzato and Müller, 2011] for proofs).

It is experienced that the MSE criterion tends to sample the function on the border of $\mathcal{D}$. To circumvent this limitation, Integrated MSE (IMSE) can be used. The aim of an IMSE design is to reduce the future integral of the MSE over $\mathcal{D}$. Thus minimising the IMSE corresponds to look for a batch of points X minimising the integral of the future MSE if the points X are added, according to the model:

$$J_{n,q}^{\mathrm{IMSE}}(X) = \int\limits_{\boldsymbol{u} \in D} c_{n,X}\,(\boldsymbol{u}, \boldsymbol{u})\,\mathrm{d}\boldsymbol{u}, \tag{2.33}$$

with $c_{n,X}(\boldsymbol{u}, \boldsymbol{u}) = \mathrm{var}\left(Y_{\boldsymbol{u}} |\, Y_{\boldsymbol{x}_1}, \ldots, Y_{\boldsymbol{x}_n}, Y_{\boldsymbol{x}_{n+1}}, \ldots, Y_{\boldsymbol{x}_{n+q}}\right)$. The term $c_{n,X}$ can theoretically be obtained using the universal kriging formula, eq. (2.8), but substantial computational saving are made using ‘update formula’, see [Chevalier et al., 2014b] for details. Figure 2.4 shows a sequential design of experiments led by the IMSE criterion. While strategies based on such criteria tend to fill the design space [Vazquez and Bect, 2011] and hence to eventually learn high-variation regions, in stationary cases it is done in a non-adaptive way as the prediction variance does not depend on observations $\boldsymbol{y}_{1:n}$ but solely on the location of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (see eq. (2.4)). In contrast, GP-based adaptive criteria have been used for estimating targeted regions according to a practical need.

2.1.3 Targeted designs

Targeted designs and Stepwise Uncertainty Reduction

There are sampling strategies that aim at evaluating as efficiently as possible f to learn a targeted feature and to answer problems such as inversion of the objective function, excursion set estimation (i.e. the set $\{x \in \mathcal{D}$ such that $f(\boldsymbol{x}) \geqslant y_0\} \subset \mathcal{D}$), probability of failure, etc. One method answering these issues is the targeted IMSE [Picheny et al., 2010]. Like the IMSE sampling criterion, the targeted IMSE computes an integral of the posterior variance over $\mathcal{D}$. The difference is that the posterior variance is multiplied by a weight function for a better exploration of the regions around the border of the excursion set. Other methods use the absolute difference between the posterior GP and the excursion threshold as in [Ranjan et al., 2008, Bichon et al., 2008], or focus on reducing the uncertainty on a volume of the excursion set [Bect et al., 2011, Chevalier et al., 2014a]. Approaches of Chevalier et al. [2013], Azzimonti [2016] estimate excursion sets by exploiting notions from random sets theory.

Some criteria can be formulated within the framework of Stepwise Uncertainty Reduction (SUR). Its aim is to provide an optimal sequence of evaluation points, selecting each point in order to reduce an uncertainty quantity. This approach requires a precise definition of an uncertainty function H_n. The function $H_n : (\mathcal{D} \times \mathbb{R})^n \to \mathbb{R}^+$ gives the remaining uncertainty after n evaluations $y_i = Y_{\boldsymbol{x}_i}$, with $(\boldsymbol{x}_i, y_i)_{i=1,\ldots,n} \in (\mathcal{D} \times \mathbb{R})^n$.

Because the design of experiments cannot be known before evaluating the random process Y, the evaluations points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ are considered as realisations of random vectors in $\mathcal{D}$. (see e.g. [Bect et al., 2011, González et al., 2016] on practical designs of experiments using SUR). However, as it does not consider a distribution on $\mathcal{D}$, the one-step-lookahead simplification is the most straightforward SUR way to get tractable sampling criteria. With an appropriate choice of uncertainty function H_{n+1}, it encompasses several sampling criteria as, among others, (multipoint) EI [Ginsbourger and Le Riche, 2010] and IMSE. The idea is to select the next evaluation by minimising the expected

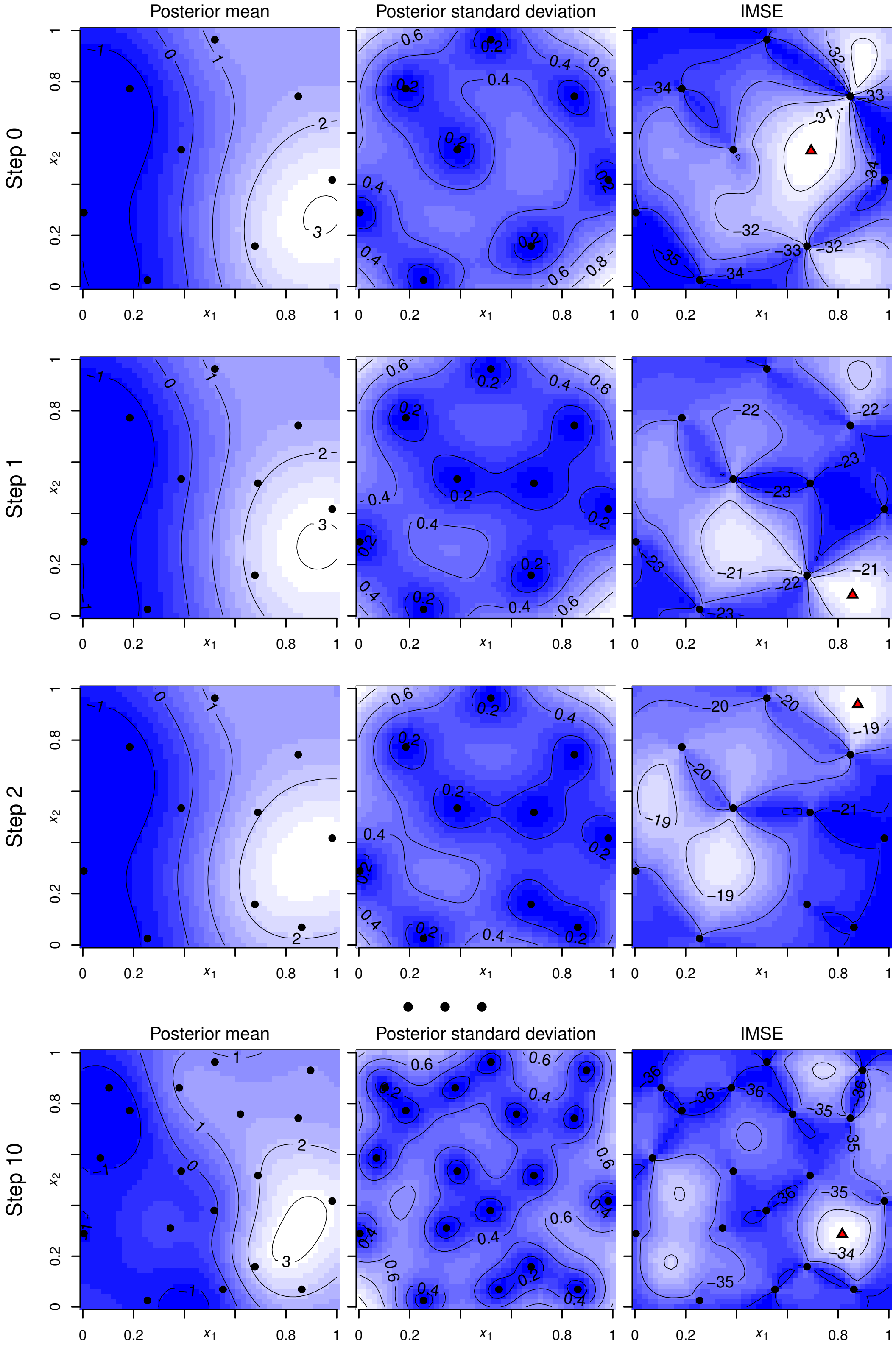


Figure 2.4: Sequential design of experiments with IMSE criterion ($n_0 = 8$, $q = 1$, next evaluation points are represented by red triangles).

uncertainty at the next step, i.e.,

$$\boldsymbol{x}_{n+1} \in \underset{\boldsymbol{x}\in\mathcal{D}}{\operatorname{argmin}}\ \mathbb{E}\left(H_{n+1}\left(\left(\boldsymbol{x}_1, y_1\right), \ldots, \left(\boldsymbol{x}_n, y_n\right), \left(\boldsymbol{x}, Y_{\boldsymbol{x}}\right)\right)\right). \tag{2.34}$$

More generally for multipoint sampling, we define

$$\begin{aligned} &X_{n+1:n+q} \in \underset{\breve{x}_{n+1},\ldots,\breve{x}_{n+q}\in\mathcal{D}}{\operatorname{argmin}} \\ &\mathbb{E}\left(H_{n+q}\left(\left(\boldsymbol{x}_1, y_1\right), \ldots, \left(\boldsymbol{x}_n, y_n\right), \left(\breve{\boldsymbol{x}}_{n+1}, Y_{\breve{\boldsymbol{x}}_{n+1}}\right), \ldots, \left(\breve{\boldsymbol{x}}_{n+q}, Y_{\breve{\boldsymbol{x}}_{n+q}}\right)\right)\right). \end{aligned} \tag{2.35}$$

Some criteria developed in chapter 4 adopt this framework of one-step-lookahead SUR. A more complete introduction on SUR strategies is given in appendix B. For more details, see e.g. [Bect et al., 2016].

Expected improvement criterion

Adaptive design criteria have also been used for derivative-free global minimisation of expensive to evaluate functions. While a number of criteria have been proposed in the literature [O'Hagan, 1978, Sacks et al., 1989, Schonlau, 1997, Jones et al., 1998, Osborne, 2010, Srinivas et al., 2010, Snoek et al., 2012, Jones, 2001, Frazier et al., 2008, Villemonteix et al., 2009, Srinivas et al., 2010, Picheny et al., 2013, Contal et al., 2014, Binois et al., 2015] and references therein, we concentrate here on the *Expected Improvement* (EI) criterion [Mockus, 1989, Jones et al., 1998], and its use in batch-sequential optimisation. The expected improvement (EI) criterion and its multipoint version are notable criteria that have an easy interpretation. Since part of this thesis concerns EI criteria, a detailed introduction is given in appendix A.

Definition of multipoint expected improvement. Denoting by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{D}$ points where f is assumed evaluated and by $\boldsymbol{x}_{n+1:n+q} := (\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q}) \in \mathcal{D}^q$ a batch of candidate points where to evaluate f next, the multipoint EI of batchsize q (or for short q-EI) is defined as

$$\mathrm{EI}_{n,q}(\boldsymbol{x}_{n+1:n+q}) = \mathbb{E}_n\left(\left(\min_{i=1,\ldots,n} Y_{\boldsymbol{x}_i} - \min_{j=n+1,\ldots,n+q} Y_{\boldsymbol{x}_j}\right)_+\right), \tag{2.36}$$

where $\mathbb{E}_n$ refers to the conditional expectation knowing the event $\mathcal{A}_n := \{Y_{\boldsymbol{x}_1} = f(\boldsymbol{x}_1), \ldots, Y_{\boldsymbol{x}_n} = f(\boldsymbol{x}_n)\}$. One way of calculating such criterion is to rely on Monte Carlo simulations. However, working on closed form formulas is a key for efficiently optimising q-EI.

Analytical derivation of expected improvement. For $q = 1$, it is well known that EI can be expressed in closed form as a function of the posterior mean and variance m_n and $\sigma_n : \boldsymbol{x} \to \sqrt{c_n(\boldsymbol{x}, \boldsymbol{x})}$. The calculation happens to involve a first order moment of the truncated univariate Gaussian distribution (see eq. (2.36) in appendix A). As shown in [Chevalier and Ginsbourger, 2014.], it turns out that eq. (2.36) can be expanded in a similar way in the multipoint case ($q \geqslant 2$) relying on moments of truncated Gaussian vectors. Given its practical importance, the question of parallelising EI algorithms and alike by selecting $q > 1$ points per iteration has been already tackled in a number of works including notably [Queipo et al., 2006, Taddy et al., 2009, Janusevskis et al., 2012, Frazier, 2012, Contal et al., 2013]. In this thesis we essentially focus in section 4.2 on approaches relying on the maximisation of eq. (2.36) and related multipoint criteria, notably by deriving closed-form formulas and fast approximates in section 4.2.

2.2 Focus on non-stationary GP modelling

2.2.1 Overview

Non-stationary GP models allow the injection of prior knowledge about spatial-dependency. Let us start by presenting several ways to produce non-stationary covariances. These methods often use a given covariance function, say k, often stationary, to create a non-stationary covariance c. Beside being a valid covariance function on $\mathcal{D}$, no further requirement is needed on k.

Vertical scaling. A first way to produce non-stationary covariance is vertical (or output) scaling [MacKay, 1998]. A stationary covariance k necessarily have a constant variance $\sigma^2 = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{x} \in D$ (say σ positive). One can then create a class of non-stationary covariance c by making this variance space dependant. More precisely, for $\boldsymbol{x} \in \mathcal{D}$

$$c(\boldsymbol{x}, \boldsymbol{x}') = \frac{\sigma(\boldsymbol{x})\sigma(\boldsymbol{x}')}{\sigma^2} k(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.37}$$

where $\boldsymbol{x} \to \sigma(\boldsymbol{x})$ is a non-negative (and non-constant) function on $\mathcal{D}$.

Composite Gaussian process. In a composite Gaussian process (CGP) model [Ba et al., 2012], the prior covariance c is a sum of two of two covariances:

$$c(\boldsymbol{x}, \boldsymbol{x}') = k_{\text{global}}(\boldsymbol{x}, \boldsymbol{x}') + \sigma(\boldsymbol{x})\sigma(\boldsymbol{x}')k_{\text{local}}(\boldsymbol{x}, \boldsymbol{x}') \tag{2.38}$$

The covariance functions k_{global} and k_{local} are typically stationary. The idea is to capture the trend of the data with k_{global}, required to be smoother, meaning e.g. that it has long correlation length, and k_{local} models the local details. As vertical scaling, the non-negative function σ produce the non-stationarity of c. Estimation of this model is not detailed here (see [Ba et al., 2012]).

Convolution methods. Some non-stationary covariances can be obtained by (spatial) convolution methods [Higdon, 2002, Gibbs, 1997]. Such a function is defined as:

$$c(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{R}^{d'}} g_{\boldsymbol{x}}(\boldsymbol{u}) g_{\boldsymbol{x}'}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} \tag{2.39}$$

where $(g_{\boldsymbol{x}})_{\boldsymbol{x}\in\mathcal{D}}$ is a family of integrable functions on $\mathbb{R}^{d'}$, $d' \in \mathbb{N}^\star$. There is no particular restriction on $g_{\boldsymbol{x}}$. For example, they can take negative values. See [Paciorek [2003], p. 26] for a proof of the definite positiveness of c. The key of this method is to obtain an analytical formula for c. It was initially derived for squared exponential (or Gaussian) kernels $g_{\boldsymbol{x}} : \boldsymbol{t} \to \varphi_{d,\Sigma_{\boldsymbol{x}}}(\boldsymbol{t} - \boldsymbol{x})$, with $d' = d$ and $(\Sigma_{\boldsymbol{x}})_{\boldsymbol{x}\in\mathcal{D}}$ a family of positive definite matrices of size d. The obtained non-stationary covariance is:

$$\begin{aligned} c(\boldsymbol{x}, \boldsymbol{x}') &= \varphi_{d,\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{x}'}}(\boldsymbol{x} - \boldsymbol{x}') \\ &= \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_{\boldsymbol{x}} + \Sigma_{\boldsymbol{x}'})}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^\top (\Sigma_{\boldsymbol{x}} + \Sigma_{\boldsymbol{x}'})^{-1} (\boldsymbol{x} - \boldsymbol{x}')\right) \end{aligned} \tag{2.40}$$

We see that this covariance is closely related to the anisotropic stationary covariance as we observe in the exponential the squared Mahalanobis distance $(\boldsymbol{x} - \boldsymbol{x}')^\top\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{x}')$, but with a location-dependent positive definite matrix $\Sigma = \Sigma_{\boldsymbol{x}} + \Sigma_{\boldsymbol{x}'}$. Although the non-stationary effects are partially due to a vertical scaling, via the space-dependant variance $c(\boldsymbol{x}, \boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d 2^d \det \Sigma_{\boldsymbol{x}}}}$, it is clear that this covariance is different from a simple vertical-scaling covariance. Indeed, if we divide by the standard deviations (or force $\det \Sigma_{\boldsymbol{x}}$ to be constant on $\mathcal{D}$) the obtained correlation function stays non-stationary. In particular on can change the anisotropy by controlling the eigen decomposition of $\Sigma_{\boldsymbol{x}}$ in different regions of $\mathcal{D}$. Paciorek [2003], p. 30, provides a way to extend this method to any covariance structure, showing that replacing the square exponential function in eq. (2.40) to any isotropic correlation function positive

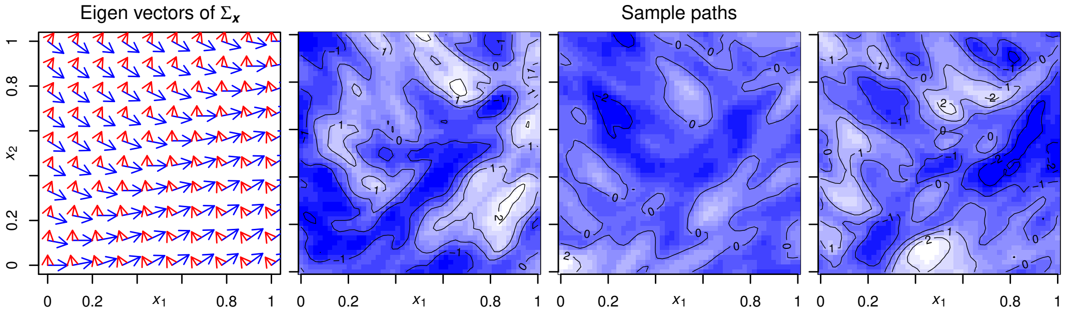


Figure 2.5: Illustration of a non-stationary GP with the convolution method. Each pair of arrows show the directions of the eigenvectors of $\Sigma_{\boldsymbol{x}}$. Their lengths are proportional to the corresponding eigenvalues.

definite on $\mathbb{R}^d$ keeps a valid (non-stationary) covariance. Figure 2.5 shows an example of this convolution method with a Matérn structure, $\nu = 5/2$.

2.2.2 Input space warping

Warping stationary GPs for creating non-stationary GPs is a common method (see, e.g., [Sampson and Guttorp, 1992]). In this approach, sometimes called the non-linear map method, the non-stationary covariance function c is obtained by chaining the stationary covariance with a warping $\boldsymbol{\gamma}$ of the input space[7].

Definition 4 (Warped stationary Gaussian process)**.** *Given a set $\mathcal{E}$, a stationary GP $Z = (Z_{\boldsymbol{x}})_{\boldsymbol{x} \in \mathcal{E}}$ and a function $\boldsymbol{\gamma} : \mathcal{D} \to \mathcal{E}$, we define the warped stationary GP associated with Z and $\boldsymbol{\gamma}$ as the process Y indexed by $\mathcal{D}$ and characterised by*

$$\forall \boldsymbol{x} \in \mathcal{D}, Y_{\boldsymbol{x}} = Z_{\gamma(\boldsymbol{x})}. \tag{2.41}$$

Without any restriction on $\boldsymbol{\gamma}$, Y is a GP and its mean and covariance functions are given by $m(\boldsymbol{x}) = \mu(\boldsymbol{\gamma}(\boldsymbol{x}))$ and $c(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{\gamma}(\boldsymbol{x}), \boldsymbol{\gamma}(\boldsymbol{x}'))$, where μ and k are the mean and covariance functions of Z, respectively.

[7] In the literature, the words 'warping' and 'deformation' are used interchangeably to describe distortion of an object or an image. Here we use the word 'deformation' for a diffeomorphism on a open subset of $\mathbb{R}^d$ (a differentiable bijection, with differentiable inverse) and the word 'warping' for a chaining function, meaning that there is not necessarily bijectivity nor same dimensionality between image and preimage.

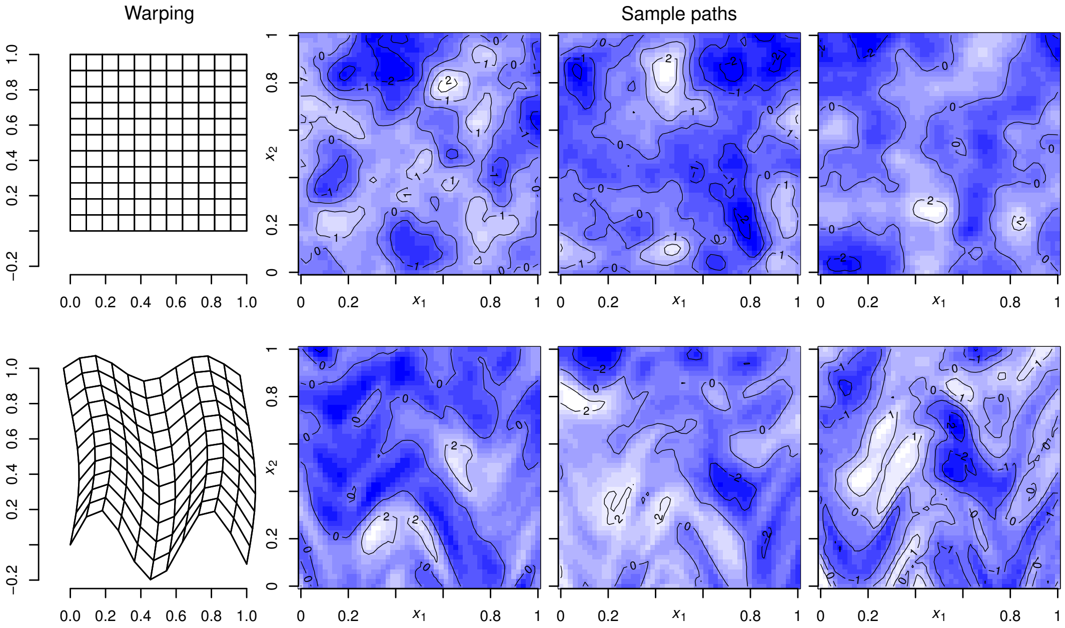


Figure 2.6: Effect of an input space warping. The warping is represented by its effect on a grid of $[0,1]^2$. The first line correspond to an isotropic GP (no warping). The second line corresponds to the warping of eq. (2.42)

.

The warping $\boldsymbol{\gamma}$ can be considered as a function that maps the data in another space in order to efficiently use a simpler model with covariance k. As k is often stationary, applying $\boldsymbol{\gamma}$ to points of a data set is sometimes called 'stationarisation'. Ideally, the image of $\boldsymbol{\gamma}$ should be a latent space in which variations of the modelled function f are smooth. Such warping corresponds to what is often referred to a change of time in stochastic process theory: from a stationary GP Z of covariance k and see $\boldsymbol{\gamma}$ as a change of its coordinate in order to create a non-stationary GP $Y_{\boldsymbol{x}} = Z_{\boldsymbol{\gamma}(\boldsymbol{x})}$. The dimension of k, $p \in \mathbb{N}^\star$, is equal to the output dimension of $\boldsymbol{\gamma}$, which is not necessarily d.

Figure 2.6 shows a warping of a stationary GP and how it impacts its sample paths. An arbitrary warping is defined for the illustration:

$$\boldsymbol{\gamma} : \boldsymbol{x} \to \left(x_1 + \frac{\sin(4x_2)}{20}, x_2 + e^{-x_2} \frac{\sin(10x_1)}{5} \right)^\top . \tag{2.42}$$

We see that the reshaping of the input space affects the sample paths. For example, when the input space is locally dilated by $\boldsymbol{\gamma}$, more variations are observed.

The estimation of the unknown warping $\boldsymbol{\gamma}$ is crucial. It is a difficult problem as possible warpings lay in a space of functions from $\mathcal{D}$ to $\mathbb{R}^p$ (with p potentially unknown). A natural idea is to consider $\boldsymbol{\gamma}$ as a deterministic parametric function to estimate. For example, Calandra et al. [2016] defines $\boldsymbol{\gamma}$ as a multi-layer neural networks. A consistent approach for estimating warpings of a bivariate GP from dense evaluations of a single (warped) realisation is provided by [Anderes and Stein, 2008]. In contrast, we consider here warping estimation from scarce evaluations in order to build appropriate non-stationary models for functions with arbitrary d-dimensional input space. In the following paragraphs, we review some warping approaches in the field of (expensive) computer experiments.

Example of a parametrised univariate warping. As an example of univariate warping, the beta cumulative distribution function is defined for $\rho_1, \rho_2 > 0$ as

$$I_{\rho_1,\rho_2} : x \to \begin{cases} \frac{B(x;\rho_1,\rho_2)}{B(1;\rho_1,\rho_2)} & \text{if} \quad x \in [0,1] \\ 0 & \text{if} \quad x < 0 \\ 1 & \text{if} \quad x > 1 \end{cases} \tag{2.43}$$

with $B(\cdot;\rho_1,\rho_2)$ is the incomplete beta function defined on $[0,1]$:

$$B(\cdot,\rho_1,\rho_2) : x \to \int_0^x t^{\rho_1-1}(1-t)^{\rho_2-1}\mathrm{d}t. \tag{2.44}$$

Note that B is linked with the gamma function:

$$B(1,\rho_1,\rho_2) = \frac{\Gamma(\rho_1)\Gamma(\rho_2)}{\Gamma(\rho_1+\rho_2)}. \tag{2.45}$$

The choice of the function I_{ρ_1,ρ_2} as an univariate warping is motivated by the wide range of warping shapes possible with only two parameters $\rho_1,\ \rho_2$ (see fig. 2.7). This function has been used e.g. in [Snoek et al., 2014] for defining univariate deformations.

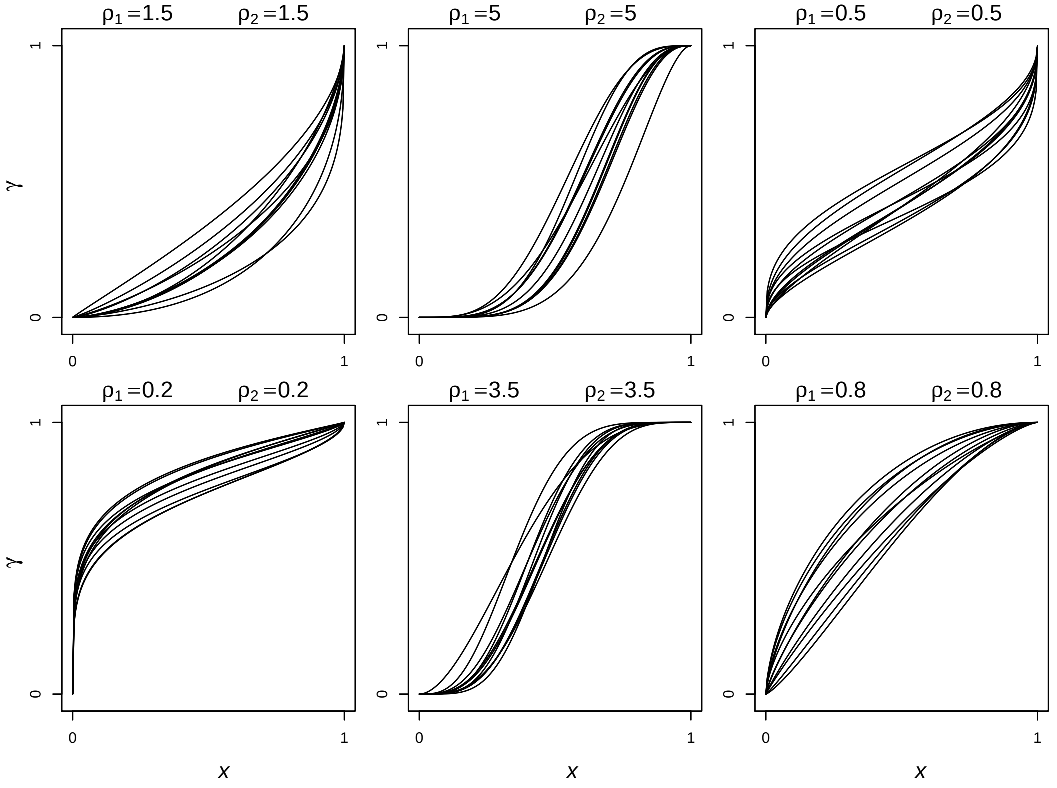


Figure 2.7: Different cumulative beta distribution functions I_{ρ_1,ρ_2}. Each panel corresponds to parameters values around given $\rho_1,\ \rho_2$.

Basis function parametrisation The flexibility of the non-linear map method is challenging for the estimation of $\boldsymbol{\gamma}$ among the set of injections on $\mathcal{D}$. A first restriction is to consider only continuous injections. The estimation of $\boldsymbol{\gamma}$ is also often simplified to a finite dimensional problem taking $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{\boldsymbol{\rho}}$, with $\boldsymbol{\rho}$ a parameter vector. For example, Gibbs' method [Gibbs, 1997] formulates $\boldsymbol{\gamma}_{\boldsymbol{\rho}}$ as a multidimensional line integral of non-negative density functions, that ensure its injectivity and continuity,

$$\boldsymbol{\gamma}_{\boldsymbol{\rho}}(\boldsymbol{x}) = \boldsymbol{x}_0 + \left(\int_{P_{\boldsymbol{x}}} g_i(\boldsymbol{u}) \mathrm{d}t \right)^{\top}_{i=1,\ldots,d},$$

with $P_{\boldsymbol{x}}$ a predefined curve between $\boldsymbol{x}_0$ and $\boldsymbol{x}$, for example the corresponding segment, and $\boldsymbol{u} : t \in [b_1, b_2] \to P_{\boldsymbol{x}}$, $b_1 < b_2$, is an arbitrary bijective parametrisation such that $\boldsymbol{u}(b_1)$ and $\boldsymbol{u}(b_2)$ give the endpoints of $P_{\boldsymbol{x}}$. In Gibbs' method, these density functions are expressed as linear combinations of radial basis functions. The estimation of $\boldsymbol{\gamma}$ is then reduced to the estimation of a finite number of weights.

We see in fig. 2.8 how this method allows an approximation of a given warping

$$\boldsymbol{\gamma}_0(\boldsymbol{x}) = \boldsymbol{x} + 1/10 \ \arctan(30(x_1^2 + x_2 - 1)). \tag{2.46}$$

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.

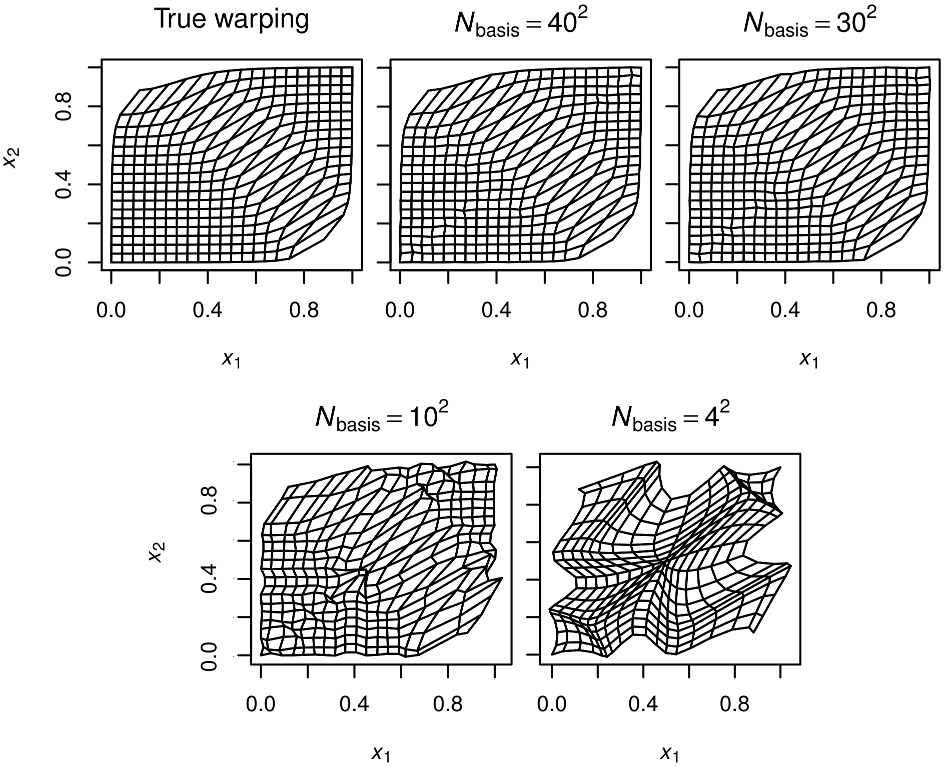


Figure 2.8: Warping approximation with Gibbs' method, for different number of basis functions. An arbitrary warping (equation (2.46)), is represented on the left by the deformation of the grid $(\frac{i}{18}, \frac{j}{18})_{i,j=0,\ldots,18}$. Then we display its approximations with different levels of precision.

We observe a degradation of the warping approximation with decreasing numbers of parameters: 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 limited 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 [Xiong et al., 2007].

Axial warping simplification In this method, it is assumed that for $\boldsymbol{x} \in \mathcal{D}$, $\boldsymbol{\gamma}(\boldsymbol{x}) = (\boldsymbol{\gamma}_i(x_i))^\top_{i=1,\dots,p}$, with $(\boldsymbol{\gamma}_i)_{i=1,\dots,p}$ continuous univariate bijections, $p \in \mathbb{N}^\star$ (p is not necessarily equal to d). Thus we have for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$

$$c(\boldsymbol{x}, \boldsymbol{x}') = k\left((\boldsymbol{\gamma}_i(x_i))^\top_{i=1,\dots,p}, (\boldsymbol{\gamma}_i(x_i'))^\top_{i=1,\dots,p}\right) \tag{2.47}$$

The axial warpings $\boldsymbol{\gamma}_i$, $i = 1, \dots, p$, are taken as piecewise second degree polynomials, with differentiability constraints and nodes placed along the i^{th} dimension. In fig. 2.9 we display the results of applying this method to the toy function

$$f : \boldsymbol{x} \in [0,1]^2 \to \frac{\sin(15x_1) + \cos(10x_2)}{5} + \arctan\left(\frac{20(x_1 + x_2) - 15}{2}\right). \tag{2.48}$$

In some situations, warping only along canonical axis can be questioned. For instance, if the expected, or 'real', warping is of the form $\boldsymbol{\gamma}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\gamma}_1(\boldsymbol{x}^\top \boldsymbol{u})\boldsymbol{u}$, 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 $\boldsymbol{\gamma}$. In Xia [2008] the number of parameters is reduced but this simplification appears to be too rigid in some applications.

2.2.3 Treed Gaussian processes

Another strategy for functions with high variation zones, which is closely related to GP models without being exactly one, is the (Bayesian) Treed Gaussian Process (TGP, [Gramacy and Lee, 2008]). It is based on partitioning the input space. Different GP models are then constructed independently in each

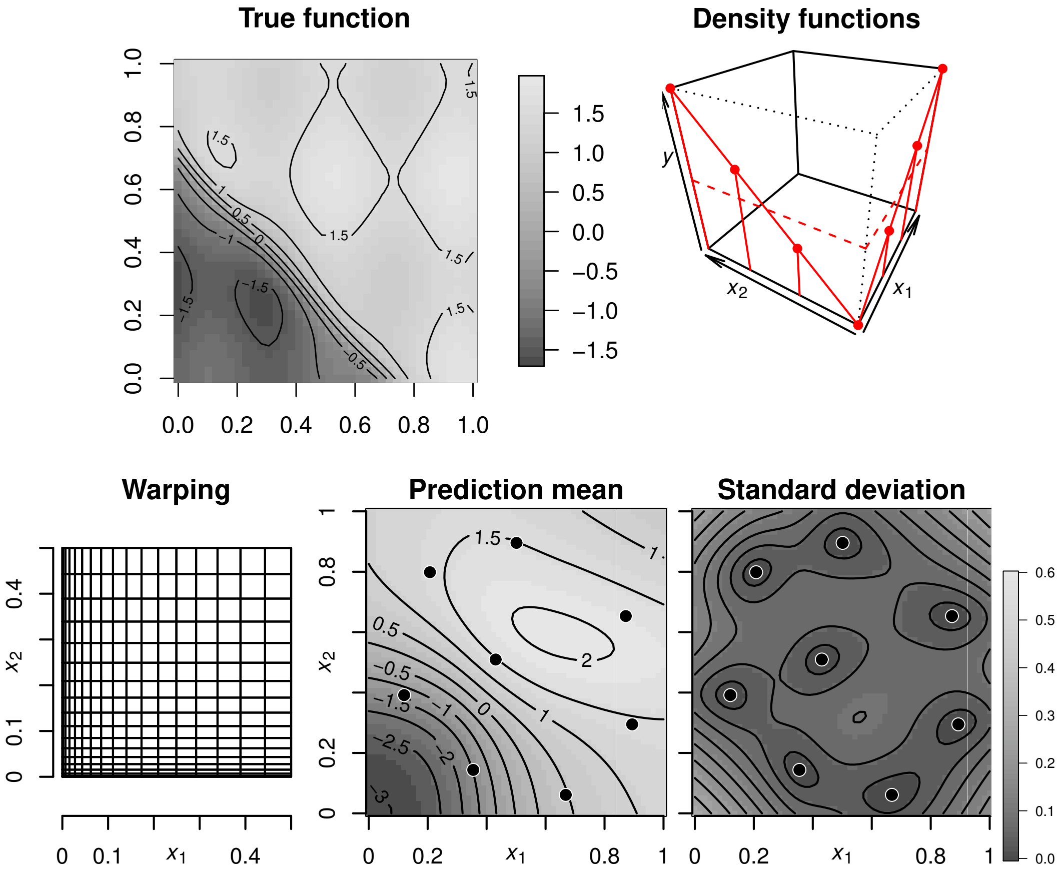


Figure 2.9: Gaussian process model with axial warping, applied to the example function (top left, eq. (2.48)). Top right: estimated warping density functions for the axes; bottom: corresponding surface warping of $[0, 1]^2$ (represented by deformation of a 10 × 10 regular orthogonal grid); prediction mean and standard deviation.

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. This extends the partitioning ideas of [Chipman et al., 1998] from simple Bayesian constant or linear models to general independent GP models.

For fixed the partitions $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_M\}$, $M \in \mathbb{N}^\star$, with, for each partition, a known covariance parameter vector $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}\} = \boldsymbol{\theta}$, the prediction from the data $\mathcal{A}_n$ can be written as

$$\mathbb{E}\left(Y_{\boldsymbol{x}}|\mathcal{A}_n, \boldsymbol{\theta}, \mathcal{P}\right) = \sum_{i=1}^{M} \mathbb{E}\left(Y_{\boldsymbol{x}}|\mathcal{A}_n^{(i)}, \boldsymbol{\theta}^{(i)}\right) \mathbb{1}_{\boldsymbol{x} \in \mathcal{P}_i} \tag{2.49}$$

with $\mathcal{A}_n^{(i)}$ the evaluations in partition $\mathcal{P}^{(i)}$. The discontinuity resulting from partitioning could sound like a drawback, although it is shown that it does not increase notably the prediction error as the model can capture smoothness through Bayesian averaging. Indeed a prior distribution for all possible partitioning $\mathcal{P}$ is fixed using a treed partitioning method [Chipman et al., 1998]. It is a recursive method, in the sense that the overall number of partitions increases by creating new sub-partitioning of existing partitions (new branches), the leafs of the tree corresponding to the overall partitions. The prediction (eq. (2.49) and eq. (2.4)) is averaged out by integrating over possible trees and covariance parameters, using sampling methods (Markov chain Monte Carlo [Richardson and Green, 1997]). Typically, the prior distribution on $\mathcal{P}$ is such that a partitioning is more likely to be sampled if it is simpler (less deep tree with less leaves) and can still explain the heterogeneity of the data set.

Figure 2.10 shows the application on the toy function obtained with the R package 'tgp' [Gramacy, 2007], and [Gramacy and Taddy, 2010].

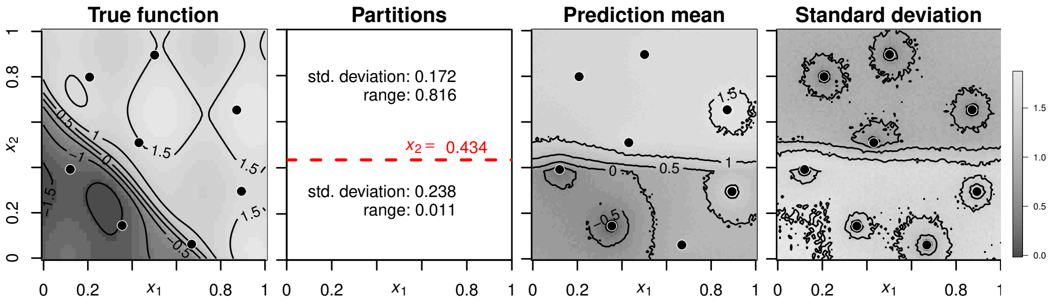


Figure 2.10: Bayesian treed Gaussian process model. **From left to right:** the objective function with an initial design; a sketch of the input space partition, the red line divides the space in two regions with different range and standard deviation for different GP models, the prediction mean and 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 $x_2 \leqslant 0.465$. The partitions here are implemented to be defined in terms of the canonical axes. We can expect that applying the method after rotating the data (following a principal component analysis) or using a single-index-based correlation structure as implemented in the package could improve the modelling.

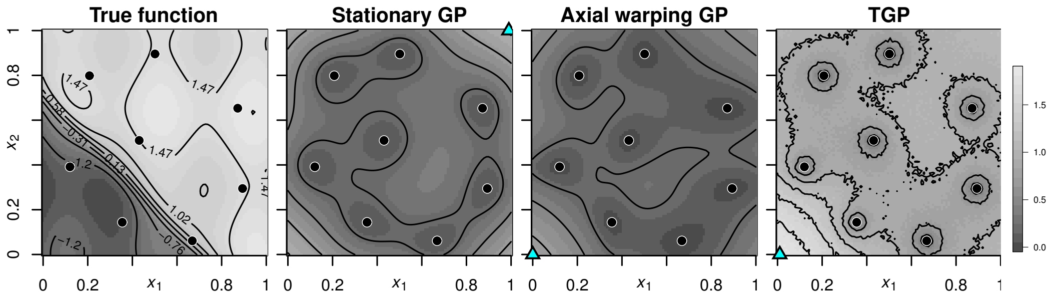


Figure 2.11: Different GP models of a function $f : \boldsymbol{x} \in [0,1]^2 \rightarrow \frac{\sin(15x_1)+\cos(10x_2)}{5} + \arctan\left(\frac{20(x_1+x_2)-15}{2}\right)$. The different models are stationary anisotropic, Xiong's axial approach, and treed GP. For each method, we see the first step of the sequential design of experiments, displaying the MSE criterion and the selected point for the next evaluation (blue triangle).

To conclude this section on non-stationary modeling, fig. 2.11 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 models (stationary, non-stationary with axial warping and TGP). The proposals for next evaluations, obtained from three MSE maxima, depend on the model for the running example function.

2.2.4 A detour through scale analysis

Before presenting in chapter 3 our results on a non-parametric approach to warping specification and estimation, let us here focus on the interplay between local analysis and warping, both in deterministic and stochastic settings. We will consider situations where $d = 1$ and $\mathcal{D} = \mathcal{E} = [0, 1]$ and where $\boldsymbol{\gamma}$ is continuous and increasing over $[0, 1]$.

One of the advantages of the wavelet transform is the possibility of simply representing in the wavelet space the actions of some operators. Denote by D_γ the composition operator associated with γ, then we obtain, for instance with the affine warping $\gamma(x) = ax + b$ ($(a, b) \in \mathbb{R}_+^\star \times \mathbb{R}$):

$$\forall(\tau, s) \in \mathbb{R}^2,\ \mathcal{W}_{D_\gamma Z}(\tau, s) = \frac{1}{\sqrt{a}}\mathcal{W}_Z(a\tau + b, s + \log_q(a)), \tag{2.50}$$

where $\log_q$ stands for the base-q logarithm. If γ is differentiable, it can be formulated in the neighbourhood of τ using its tangent, $\gamma(x) = \gamma(\tau)+\gamma'(\tau)(x-\tau) + o(|x - \tau|)$. Then we get, using Equation (2.50),

$$\mathcal{W}_{D_\gamma Z}(\tau, s) = \frac{1}{\sqrt{\gamma'(\tau)}}\mathcal{W}_Z(\gamma(\tau), s + \log_q(\gamma'(\tau))) + \varepsilon_{\mathcal{W}}(\tau, s),$$

with $\varepsilon_{\mathcal{W}}$ an error quantity. As expected, the error level depends on the quality of the local approximation of γ by its tangent. It can be proven that under fast decay assumption for the wavelet, $\varepsilon_{\mathcal{W}}$ vanishes when $s \to -\infty$[Omer and Torresani, 2016].

This important property of translation in wavelet space has been exploited in several works to estimate the warping function in the framework of signal analysis [Clerc and Mallat, 2003, Omer and Torresani, 2016]. A way to do so is to compute the local scale (as e.g. in [Flandrin, 1993]). For a deterministic function, the local scale gives the evolution of the average scale weighted by the square of the wavelet coefficients at fixed positions. For a stochastic process this definition is extended using the second order moment of the wavelet transform. More precisely, the local scale associated with a GP Y is given by:

$$G_Y(\tau) = \frac{\int_{\mathbb{R}} h^{-s}\mathbb{E}\left(|\mathcal{W}_Y(\tau, s)|^2\right) \mathrm{d}s}{\int_{\mathbb{R}} \mathbb{E}\left(|\mathcal{W}_Y(\tau, s)|^2\right) \mathrm{d}s}. \tag{2.51}$$

The following proposition shows a link between the local scale of a warped GP and its associated warping.

Proposition 4. *If $(Y_x)_{x\in\mathbb{R}}$ is stationary, G_Y is constant. Moreover, if $(Y_x)_{x\in\mathbb{R}}$ is such that $\forall x \in \mathbb{R}$, $Y_x = Z_{\gamma(x)}$ with $\gamma(x) = ax + b$, $(a, b) \in \mathbb{R}^2$, and $(Z_x)_{x\in\mathbb{R}}$ stationary, then $G_Y(\tau) = aG_Z(0)$.*

This property is common knowledge due to the simplicity of the proof. But we could not find a published proof, and for self-containedness with section 3.4, we give a proof below.

Proof. If $(Y_x)_{x\in\mathbb{R}}$ is stationary, changing the integration order for $\mathbb{E}\left(|\mathcal{W}_Y(\tau, s)|^2\right)$ gives

$$\mathbb{E}\left(|\mathcal{W}_Y(\tau, s)|^2\right) = \frac{1}{h^s}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{E}\left(Y_x Y_{x'}\right) \psi\left(\frac{x-\tau}{h^s}\right) \psi\left(\frac{x'-\tau}{h^s}\right) \mathrm{d}x\mathrm{d}x',$$

which does not depend on τ by change of variable $u = x - \tau$ and $u' = x' - \tau$, concluding the first part of the proof. As for the second assertion, assuming that $\forall x \in \mathbb{R}$, $Y_x = Z_{\gamma(x)}$ with $\gamma(x) = ax + b$ and $(Z_x)_{x\in\mathbb{R}}$ stationary, Equation (2.50) leads to:

$$\begin{aligned} G_Y(\tau) &= \frac{\int_{\mathbb{R}} h^{-s}\mathbb{E}\left(|\mathcal{W}_Z(a\tau + b, s + \log_q(a))|^2\right) \mathrm{d}s}{\int_{\mathbb{R}} \mathbb{E}\left(|\mathcal{W}_Z(a\tau + b, s + \log_q(a))|^2\right) \mathrm{d}s}, \\ &= a\frac{\int_{\mathbb{R}} h^{-u}\mathbb{E}\left(|\mathcal{W}_Z(a\tau + b, u)|^2\right) \mathrm{d}u}{\int_{\mathbb{R}} \mathbb{E}\left(|\mathcal{W}_Z(a\tau + b, u)|^2\right) \mathrm{d}u}, \\ &= aG_Z(0), \end{aligned} \tag{2.52}$$

since G_Z is constant. □

This property gives access to the slope of the affine warping. In section 3.4, we use this property for approximating the derivative of γ at a given point τ, assuming that due to the decay of the basis function, the derivative of γ is reasonably approximated by the local scale .

Chapter 3

Contributions in warped Gaussian process modelling

Let us now focus on the modelling of non-stationary GPs. The main developments concern the introduction in section 3.1 of a novel family of prior covariances for WaMI-GP based on input space warping. In sections 3.2 and 3.3, we present its main properties and we highlight its potential on a synthetic test case. In 3.4 a different approach based on warping approximation using wavelets is provided.

3.1 Formulation of the WaMI-GP model

In the context of GP modelling with expensive evaluations, an important limitation of warping methods of non-stationary covariances is the estimation of the warping function $\boldsymbol{\gamma}$. For example, we have seen in section 2.2.2 that the number of weights for parametrising $\boldsymbol{\gamma}$ as an integral of linearly combined basis functions is dN_{basis}^{d}, with N_{basis} the number of basis functions in one canonical direction. In order to adress this issue, Xiong et al. [2007] formulated $\boldsymbol{\gamma}$ as a tensor product of univariate functions $(\gamma_i)_{i=1,\ldots,d}$, for $\boldsymbol{x} \in \mathcal{D}$,

$$\boldsymbol{\gamma}(\boldsymbol{x}) = \begin{pmatrix} \gamma_1(x_1) \\ \vdots \\ \gamma_d(x_d) \end{pmatrix} \tag{3.1}$$

(see section 2.2.2 for details). Thus, the non-stationary structure is simplified by assuming that the heterogeneity in any single canonical direction does not depend on the coordinates in other canonical directions. Whereas this simplification drastically eases estimation of $\boldsymbol{\gamma}$, we would like our model to be able to detect or reproduce heterogeneity in any direction. To do so, we chain $\boldsymbol{\gamma}$ with a linear transformation $\boldsymbol{x} \to A\boldsymbol{x}$:

$$\boldsymbol{\gamma}(A\boldsymbol{x}) = \begin{pmatrix} \gamma_1(\boldsymbol{a}_1^\top \boldsymbol{x}) \\ \vdots \\ \gamma_{d'}(\boldsymbol{a}_{d'}^\top \boldsymbol{x}) \end{pmatrix}. \tag{3.2}$$

where A is a matrix with d' rows $\boldsymbol{a}_1^\top, \dots, \boldsymbol{a}_{d'}^\top \in \mathbb{R}^d$, $d' \in \mathbb{N}^\star$. We will call GP models based on this warping 'Warped Multiple Index GP model' or WaMI-GP model. These models rely on a prior covariance function computed from eq. (3.2) defined below.

Definition 5 (WaMI covariance family)**.** *Let $d' \in \mathbb{N}^\star$, $A = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_{d'}]^\top \in \mathbb{R}^{d' \times d}$, $\gamma_i(\cdot, \boldsymbol{\rho}_i) : \mathbb{R} \mapsto \mathbb{R}$ be functions parametrised by real-valued vectors $\boldsymbol{\rho}_i$ ($i = 1, \dots, d'$, $(\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_{d'}) \in \mathbb{R}^{p_1} \times \dots \times \mathbb{R}^{p_{d'}}, p_1, \dots, p_{d'} \in \mathbb{N}^\star$) and $k_{\boldsymbol{b}}$ be a positive definite kernel on $\mathbb{R}^{d'}$ parametrised by $\boldsymbol{b} \in \mathbb{R}^r$, for some $r \in \mathbb{N}^\star$. Assuming that the parametric form of the γ_i's is given and denoting by $\boldsymbol{\theta}$ a vector of parameters containing A, the $\boldsymbol{\rho}_i$'s and $\boldsymbol{b}$, we define the associated WaMI (Warped Multiple Index) kernel on $\mathcal{D}$ by*

$$c_{\boldsymbol{\theta}} : (\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{D}^2 \to c_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\boldsymbol{b}}\left(\left(\gamma_i(\boldsymbol{a}_i^\top \boldsymbol{x}; \boldsymbol{\rho}_i))\right)_{i=1,\dots,d'}, \left(\gamma_i(\boldsymbol{a}_i^\top \boldsymbol{x}'; \boldsymbol{\rho}_i)\right)_{i=1,\dots,d'} \right). \tag{3.3}$$

Two interpretation viewpoints. Let us first highlight a link between WaMI-GP models and the multi index model introduced in section 2.1.1. If a warped Y is defined as $Y = Z \circ \boldsymbol{\gamma} \circ A$ where Z is a GP on $\mathbb{R}^{d'}$ with covariance $k_{\boldsymbol{b}}$ (and arbitrary mean function), then its covariance function is given by eq. (3.3)[1]. When Y is formulated as $Y = Y^{(1)} \circ A$, with $Y^{(1)} = Z \circ \boldsymbol{\gamma}$, the equality corresponds in term of covariance to the MIM model (as in eq. (2.17)), for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$,

$$c(\boldsymbol{x}, \boldsymbol{x}') = c^{(1)}(A\boldsymbol{x}, A\boldsymbol{x}') \tag{3.4}$$

with $c^{(1)} = \operatorname{cov}\left(Z_{\boldsymbol{\gamma}(\boldsymbol{x})}, Z_{\boldsymbol{\gamma}(\boldsymbol{x}')}\right)$ the covariance function of $Y^{(1)}$. From this MIM perspective, a WaMI-GP introduces non-stationarity into the covariance by applying non-linear deformations to the result of each scalar product $\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)_{i=1,\dots,d'}$.

[1] Here A represents the linear map yield by the matrix with same notation.

From another perspective, let us assume first that A is invertible and, for all $\boldsymbol{u}, \boldsymbol{u'} \in \mathrm{Im}_A(\mathcal{D})$, that $Z^{(1)}_{\boldsymbol{u}} = Y_{A^{-1}\boldsymbol{u}}$. Then we get the formulation of the axial warping model (as in eq. (2.47)),

$$k^{(1)}(\boldsymbol{u}, \boldsymbol{u'}) = k\left((\gamma_i(u_i))^\top_{i=1,\dots,d'}, (\gamma_i(u'_i))^\top_{i=1,\dots,d'}\right) \tag{3.5}$$

with $k^{(1)}$ the covariance function of $Z^{(1)}$. In this set-up WaMI-GP models allow non-canonical directions for orientation of the univariate deformations by acting on the input space via a linear map with matrix A.

In summary the WaMI-GP family of models corresponds to an extension of two formulations: it combines axial deformations (subcase $A = \mathrm{Id}_n$) and multiple index modelling (subcase $\gamma_i : x \to x$ for all $i = 1, \dots, d'$).

Reduction or inflation of dimension. The covariance kernel introduced in eq. (3.3) accomodates dimension reduction (thus reducing the number of axial warpings) when $d' < d$. The rectangular matrix $A \in \mathbb{R}^{d' \times d}$ maps the input space to an image space of lower dimension. As we will see in what follows, the reverse case $d' > d$ is useful for modelling function f with more complex spatial heterogeneity.

It is possible to take identity γ_i's for one to several dimensions, hence reducing the number of deformations and covariance parameters. Besides this, the class can be generalised to cases where the warpings $(\gamma_j)_{j=1,\dots,r}$ are not univariate but rather defined on subspaces $\left(\mathbb{R}^{d'_j}\right)_{j=1,\dots,r}$, with $r \leqslant d'$, $d'_1, \dots, d'_r \in \mathbb{N}^\star$ and $\sum\limits_{j=1}^{r} d'_j = d'$. This extension requires parametrisation of multivariate warpings and is not developed in this thesis.

Standard parametrisation. In a general manner, the total number of parameters of the WaMI covariance is $d'd + \#\boldsymbol{b} + \sum_{i=1}^{d'} \#\boldsymbol{\rho}_i$. The implementation of a WaMI covariance from definition 5, requires choices about:

- the kernel family for $k_{\boldsymbol{b}}$,
- the family of parametrised functions for each γ_i,
- the value of d'.

In this paragraph we propose a standard parametrisation appropriate for a wide range of applications. Without any specification, these settings will be implicitly used in the rest of the thesis, although there are many other suitable choices. We choose $k_{\boldsymbol{b}}$ as a stationary, radial, Matérn kernel with $\nu = 5/2$, see equation (2.6). As discussed in section 2.1, in the absence of explicit prior knowledge on high order of differentiability, this covariance structure is commonly used in machine learning for its compromises between having short analytical formula (vs. e.g. cases of $\nu \geqslant 7/2$, $\nu < \infty$, where some formulas are still algebraic but longer), generating smooth realisations (vs. e.g. cases $\nu \leqslant 3/2$), and avoiding numerical singularities at conditioning (vs. e.g. Gaussian kernel case). Since geometric anisotropy can be seen as a result of warping a isotropic GP (see section 2.1.1), $k_{\boldsymbol{b}}$ is fixed isotropic and anisotropical effects will be included in $\boldsymbol{\gamma} \circ A$. So we write for $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$

$$k_b(\boldsymbol{x}, \boldsymbol{x}') = k^{\text{‘}5/2\text{’}} \left(\frac{||\boldsymbol{x} - \boldsymbol{x}'||}{b} \right). \tag{3.6}$$

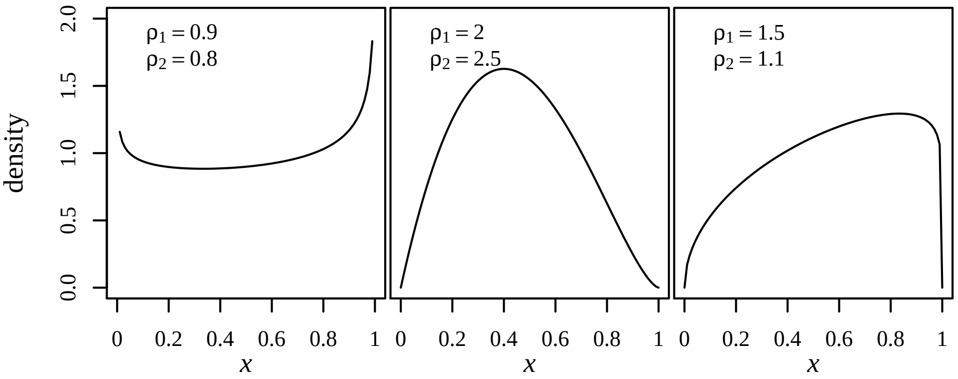


Figure 3.1: Example of warping densities taken from the beta distribution $I'_{\rho_{i,1},\rho_{i,2}}$

For the choice of γ_i's, we propose to use a class of transformations: the cumulative distribution functions of beta distributions $I_{\rho_{i,1},\rho_{i,2}}$, $\rho_{i,1}, \rho_{i,2} > 0$ see section 2.2.2. This class combines practical properties, such as bijectivity and differentiability, is capable of expressing a fair range of warping shapes and provides a lean parametrisation with only two shape parameters. When it is not U-shaped, the density of the beta distribution is unimodal meaning that the warpings of the axes are either successive contraction-dilatation-contraction or dilatation-contraction-dilatation of three partitioning intervals (see fig. 3.1). Moreover in cases of a middle dilatation (higher density inside $]0, 1[$, i.e. $\rho_{i,1}$, $\rho_{i,2} > 1$), the density goes to zero at endpoints 0 and 1. This very low value implies a very strong contraction (low density) of $[0, 1]$ at its endpoints, no matter the value of the parameters ρ_1, $\rho_2 > 1$. To relax this strong assumption, the warping is combined with a linear function:

$$\gamma_i(\cdot;\boldsymbol{\rho}_i) : x \to \frac{1}{1+\rho_{i,3}}\left(\rho_{i,3}\left(x+\frac{1}{2}\right)+I_{\rho_{i,1},\rho_{i,2}}\left(x+\frac{1}{2}\right)\right)-\frac{1}{2} \tag{3.7}$$

with $\rho_{i,3} \geqslant 0$. The shifts of $1/2$ and the normalisation term are for conserving the same image of $\gamma_i([-1/2,1/2];\boldsymbol{\rho}_i) = [-1/2,1/2]$. The new parameter $\rho_{i,3}$ will be often empirically set to 1. Restrictiveness of this class of warpings comes from the desire to limit the number of parameters. For large designs, it is possible to increase model flexibility, by considering more complicated parametrised warpings (as we illustrate later with a warping for modelling two high variation zones) or using basis functions for warping estimation (as in e.g. Xiong et al. [2007]).

Ideally, d' is inferred from data, as opposed to being set arbitrarily. A natural idea is to start with d' low compared to d and increment it progressively while monitoring the prediction performance of the model (for example by cross validation see section 2.1.1).

A last detail of the model parametrisation is to set the origin of the linear map to a reference point $\boldsymbol{u}_0 \in \mathbb{R}^d$, i.e. replace the linear map A by a corresponding affine map $A_{\boldsymbol{u}_0} : \boldsymbol{x} \to A(\boldsymbol{x}-\boldsymbol{u}_0)$ with $\boldsymbol{u}_0 \in \mathbb{R}^d$. This insures, if $\boldsymbol{u}_0$ is in $\mathcal{D}$, that their exists a subset of $\mathcal{D}$ whose image by γ_i is included in $[-1/2,1/2]^{d'}$, the region in $\mathbb{R}^{d'}$ actually warped (an not only linearly transformed). Although the point $\boldsymbol{u}_0$ can be seen as an additional parameter, we always set for our applications $\boldsymbol{u}_0 = \frac{1}{2}\mathbf{1}_d$, the centre of $\mathcal{D} = [0,1]^d$.

Illustrations. The flexibility of the WaMI-GP as a generative model is depicted in here with various examples.

Stationary subcase. Let us first illustrate the case where all univariate deformations are the identity. In fig. 3.2 we illustrate the warped space (here the overall warping amounts to A), the WaMI kernel and corresponding GP sample paths with

$$A = \begin{pmatrix} 5 & 10 \\ 7.5 & 2.5 \end{pmatrix}. \tag{3.8}$$

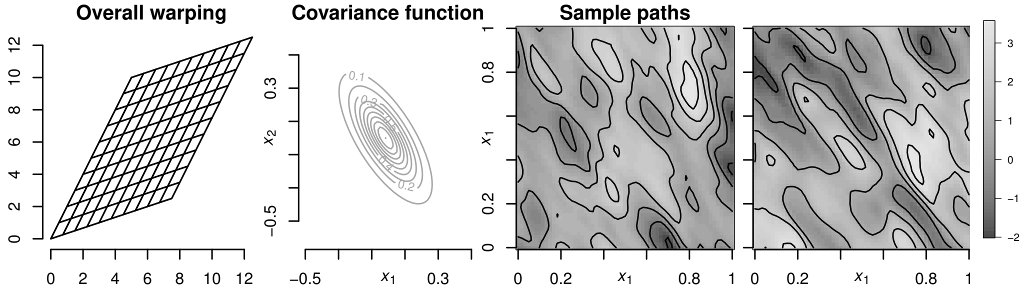


Figure 3.2: WaMI-GP model in the case of a stationary base kernel and no axial deformation. From left to right: warping represented by mapping of the grid $(\frac{i}{10}, \frac{j}{10})_{i,j=0,\dots,10}$, the covariance function $c(\cdot, (0,0)^\top)$, and two realisations sample from a centred GP with this covariance.

This case corresponds to the geometric anisotropic stationary covariance (section 2.1.1, Rasmussen and Williams [2006]). If $k_{\boldsymbol{b}}$ is an isotropic covariance on $\mathbb{R}^d$, $c_{\boldsymbol{\theta}}$ is a geometric anisotropic version with symmetric semi-definite matrix $AA^\top$. In particular, if $AA^\top$ is definite, the distance in $\mathcal{D}$ is changed to the Mahalanobis distance of matrix $AA^\top$. The first eigenvectors of $AA^\top$, ordered increasingly by their eigenvalues, 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 illustrate the case of axial deformations alone in fig. 3.3. We select A as the identity matrix, γ_1 as the identity function, i.e. $\rho_{1,1} = \rho_{1,2} = 1$, but γ_2 is non-linear with $\rho_{2,1} = \rho_{2,2} = 5$ (see eq. (3.7), with $\rho_{i,3}$ always set to 1). Except for the way to parametrise of the axial warpings, this case corresponds to eq. (3.1) in Xiong et al. [2007]. In this thesis 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.

A first non-axial warping with $d' = d$. In this example, we combine the two previous cases with a matrix rotation and one axial warping (fig. 3.4):

$$A = \frac{1}{\sqrt{2}} \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}, \text{ and } \gamma_1 : x \to x + I(x; 30, 30). \tag{3.9}$$

We observe that this covariance setting allows high variations at $x_1 + x_2 = 1$.

Example of non-canonical orientation and two high variation zones. Having a neutral parametrisation towards canonical axes is the key idea for estimating

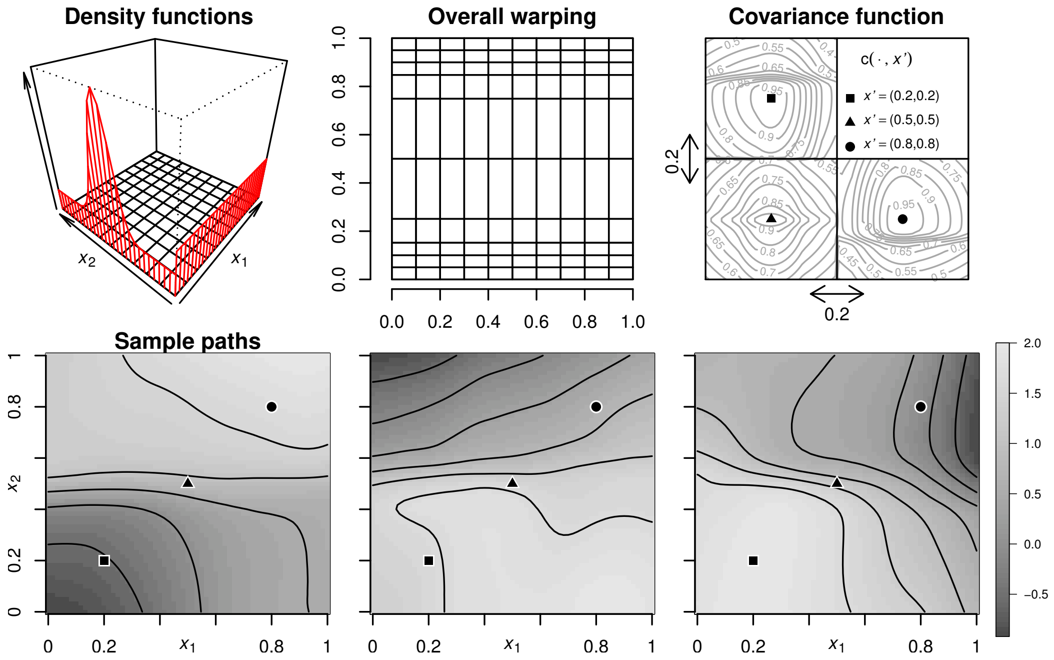


Figure 3.3: WaMI-GP with axial deformations. From left to right: density function of the deformations in each direction, warping of the grid $(\frac{i}{10}, \frac{j}{10})_{i,j=0,\ldots,10}$, the covariance function $c(\cdot, \boldsymbol{x}'))$ for different values of $\boldsymbol{x}'$, and three corresponding WaMI-GP realisations.

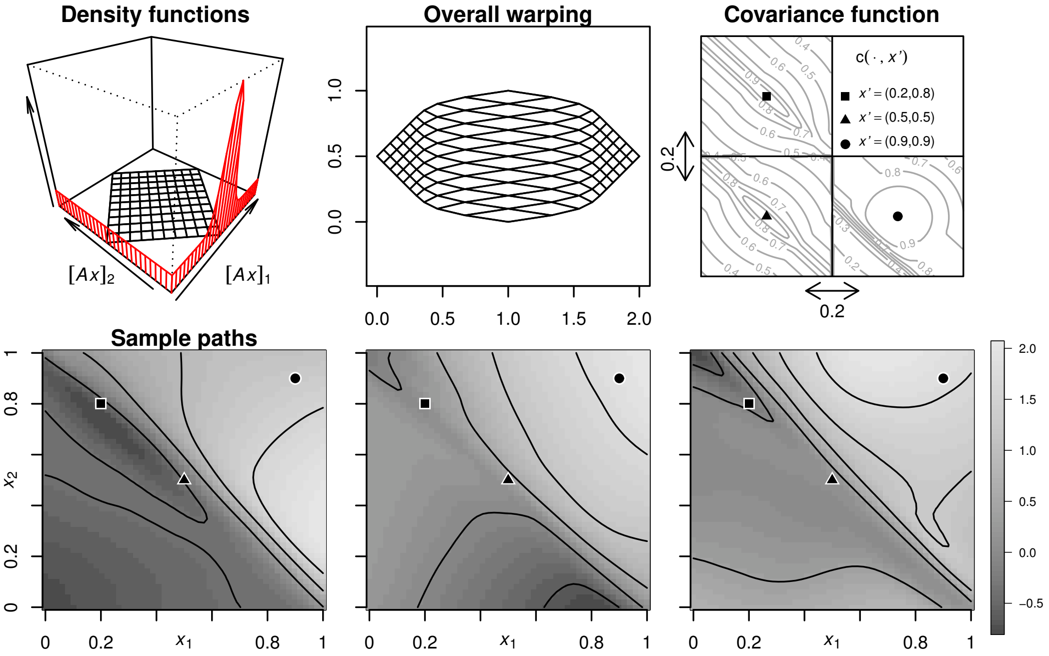


Figure 3.4: WaMI-GP with axial deformations. From left to right: density function of the deformations in each direction, warping of the grid $(\frac{i}{10}, \frac{j}{10})_{i,j=0,\ldots,10}$, the covariance function $c(\cdot, \boldsymbol{x}'))$ for different values of $\boldsymbol{x}'$, and three realisations sample from with this covariance.

arbitrary directions of heterogeneous variations. We now take

$$A = \begin{pmatrix} \cos(\pi/12) & -\sin(\pi/12) \\ \sin(\pi/12) & \cos(\pi/12) \end{pmatrix}. \tag{3.10}$$

In addition, we take here γ_2 with two ridges,

$$\gamma_2 : x \to x + I(2x; 15, 15) + I\left(2\left(x - \frac{1}{2}\right); 15, 15\right). \tag{3.11}$$

As sample paths resulting from this covariance function have two high variation zones, fig. 3.5 shows the links between GP realisations and the corresponding overall warping.

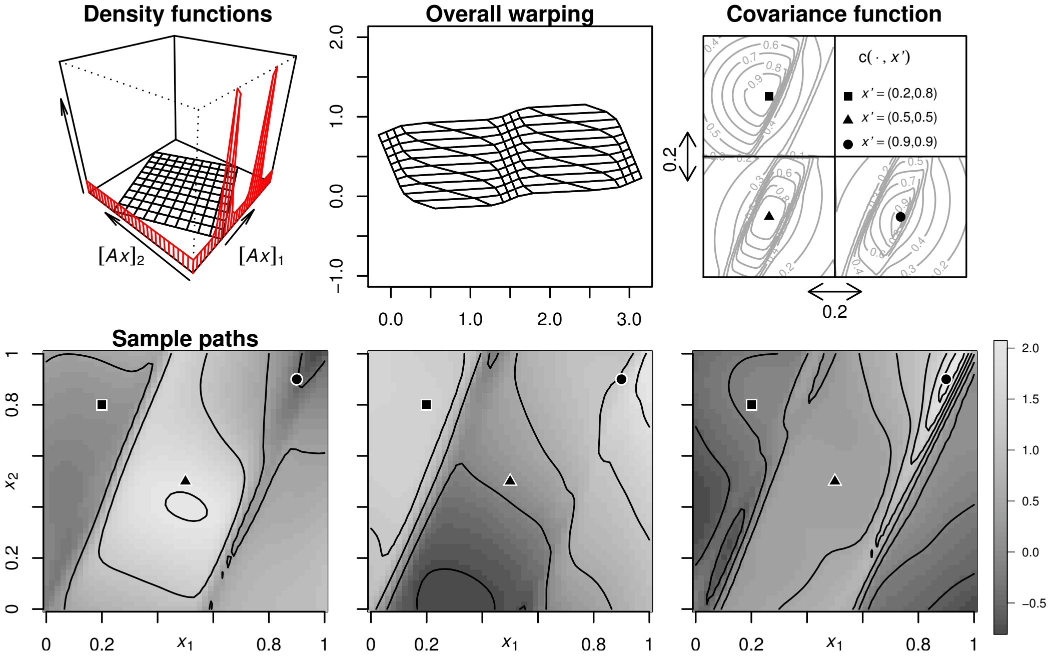


Figure 3.5: Example of a WaMI-GP with two regions of high variations. From left to right: density functions of the deformations in each direction after the linear transformation, warping of the grid $(\frac{i}{10}, \frac{j}{10})_{i,j=0,\ldots,10}$, the covariance function $c(\cdot, \boldsymbol{x}'))$ for different values of $\boldsymbol{x}'$, and three corresponding WaMI-GP realisations.

Two examples with $d' > d$. In the last two examples we experiment $d' > d$, with $d = 2$ and $d' = 3$. In fig. 3.6 we consider

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}, \tag{3.12}$$

and the warping functions

$$\gamma_1 = \gamma_2 = \gamma_3 : x \to x + I(x; 30, 30). \tag{3.13}$$

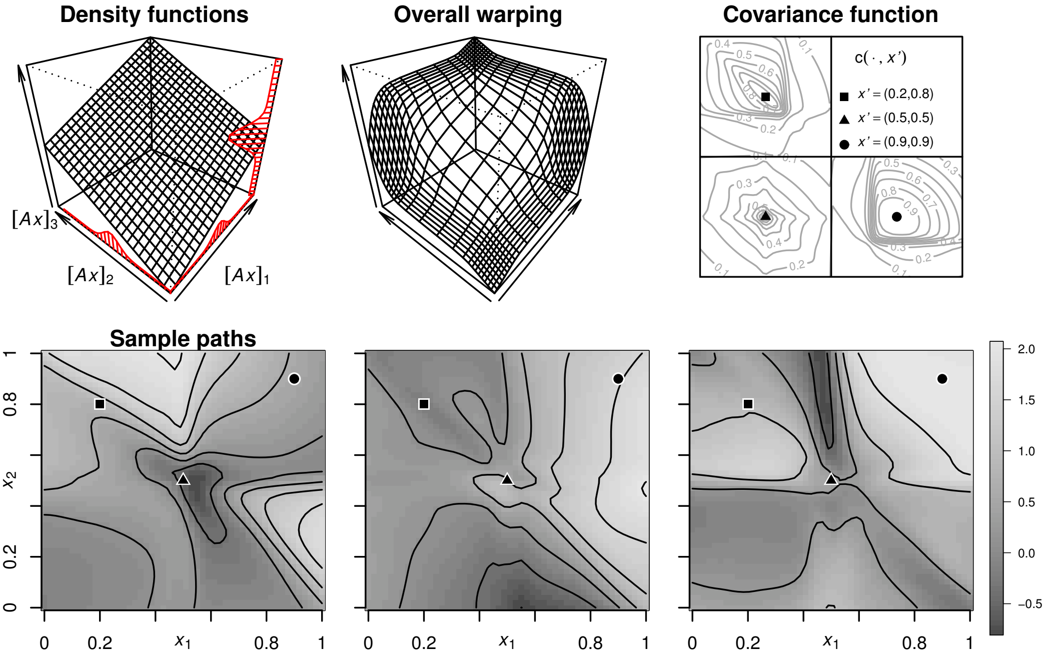


Figure 3.6: Example of a WaMI-GP with $d' > d$. High variations are concentrated in the center.

We see that the input space is warped in three dimensions. The two warpings γ_1 and respectively γ_2 dilate the space around $x_1 = 1/2$ and respectively $x_2 = 1/2$ without rotations, in the manner the previous example of axial warpings. However a third dilatation not aligned with a canonical direction, along the line $x_1 + x_2 = 0$, is added using an axial warping in the third dimension.

We give a similar example, displayed in fig. 3.7, with:

$$A = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.14}$$

and the warping functions

$$\gamma_1 : x \to x + \frac{2+\sqrt{2}}{4} + I(x + \frac{2+\sqrt{2}}{4}; 30, 30) \tag{3.15}$$

$$\gamma_2 : x \to x + \frac{2-\sqrt{2}}{4} + I(x + \frac{2-\sqrt{2}}{4}; 30, 30) \tag{3.16}$$

$$\gamma_3 : x \to x + I(x; 30, 30). \tag{3.17}$$

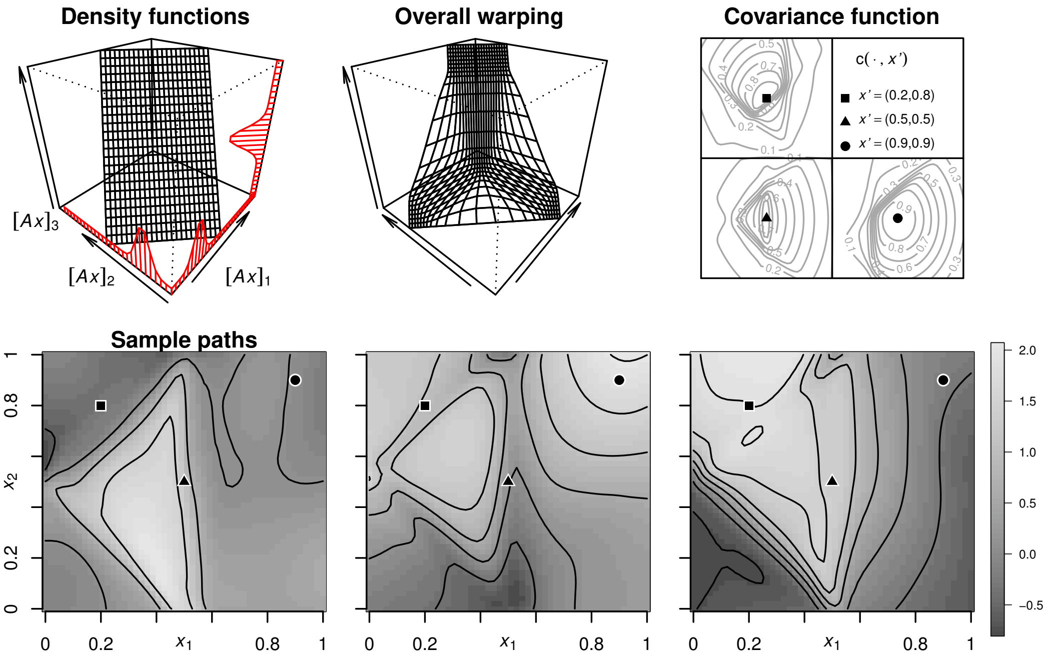


Figure 3.7: Example of a WaMI-GP with $d' > d$. We observe three directions of high variations.

This warping example in 3 dimensions creates high variations zones for the realisations organised in a triangle in the left part of the domain.

3.2 Properties of WaMI-GP

In this section, we investigate the properties of the WaMI-GP kernel and its associated (centred) WaMI-GP. We first show when the kernel is strict positive-definite. Although strict definiteness is not necessary for a covariance function, this property is useful for avoiding singularity issues with covariance matrices, in particular in the conditioning formulas of eq. (2.8), the matrix C may not be invertible without strict definiteness. Then we link the smoothness of the kernel with the differentiability of the GP. Differentiability of GP, and in particular mean-square differentiability is used for sampling strategies. For example in section 4.1 we use mean-square differentiability when sampling the heterogeneous function or in section 4.2 for global optimisation. Finally, the Jacobian determinant of the warping is calculated. The Jacobian determinant is a useful tool in the analysis of heterogeneity of the model, as it reflects the local contraction or dilatation of the input space.

3.2.1 Strict positive-definiteness

Let us first define conditions on $k_{\boldsymbol{b}}$, A, and γ_i's, under which the WaMI kernel is strictly positive definite.

Proposition 5 (Positive definiteness)**.** *Assume that $k_{\boldsymbol{b}}$ is strictly positive definite, that the $\gamma_i(\cdot;\boldsymbol{\rho}_i)$ are injective and that the rank of A is equal to d. Then the WaMI kernel of eq.* (3.3) *is strictly positive definite.*

Proof. Assuming the existence of $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{D}$, with $\boldsymbol{\gamma}(\boldsymbol{z}) = \boldsymbol{\gamma}(\boldsymbol{z}')$, gives $\forall i = 1, \dots, d'$, $\gamma_i(z_i;\boldsymbol{\rho}_i) = \gamma_i(z_i';\boldsymbol{\rho}_i)$ and thus $\boldsymbol{z} = \boldsymbol{z}'$ by injectivity of each function γ_i. Moreover, as A is full column rank, the linear map is injective. So the composition of these two maps $g = \left(\bigotimes_{i=1}^{d'} \gamma_i\right) \circ A$ ($\bigotimes$ refers to tensor product) is also injective. Finally, for all distinct $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathcal{D}$, $N \in \mathbb{N}^\star$, the points $\boldsymbol{g}_1 = \boldsymbol{g}(\boldsymbol{x}_1), \dots, \boldsymbol{g}_N = \boldsymbol{g}(\boldsymbol{x}_N)$ are distinct; therefore by exploiting the positive definiteness of $k_{\boldsymbol{b}}$ we have for all $\alpha_1, \dots, \alpha_N \in \mathbb{R}$, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathcal{D}$ distinct, $N \in \mathbb{N}^\star$,

$$\begin{aligned} \sum_{i,j=1}^{N} \alpha_i \alpha_j c_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{i,j=1}^{N} \alpha_i \alpha_j k_{\boldsymbol{b}}(g_i, g_j) \\ &= 0 \text{ if and only if } \alpha_1, \dots, \alpha_N = 0. \end{aligned} \tag{3.18}$$

□

3.2.2 Jacobian determinant

We have seen that the Jacobian (or derivatives) of the warping $\boldsymbol{\gamma} \circ A$ plays a role in the regularity of the process. Dilatation (resp. contraction) zones are areas where the warped GP has high (resp. low) variation. Here we compute the determinant of the Jacobian of the warping to analyse the location of high variation zones according to the parameters of the warping.

Proposition 6 (Jacobian of the multiple index warping)**.** *Assume that $d' = d$ and $\gamma_i(\cdot;\boldsymbol{\rho}_i)$ is differentiable on $\mathbb{R}$, for $i = 1, \dots, d$, and denote with $\gamma_i'(\cdot;\boldsymbol{\rho}_i)$ its derivative. Then for $\boldsymbol{x} \in \mathcal{D}$, the determinant of the warping $\boldsymbol{\gamma} \circ A$ is*

$$\det\left(\boldsymbol{\gamma}(A\boldsymbol{x})\right) = \det(A) \prod_{i=1}^{d} \gamma_i'\left(\boldsymbol{a}_i^\top \boldsymbol{x}; \boldsymbol{\rho}_i\right). \tag{3.19}$$

Proof. The proof is straightforward from the chain rule with Jacobian matrices and the basic properties of the determinant. □

This property shows that high variation zones are directly linked with the high univariate derivatives of γ_i. In particular, with A invertible, if γ_i is increasingly monotonic with an unique inflection point at $x_i = c_i$, $c_i \in \mathbb{R}$, for all $i = 1, \ldots, d$, the point of highest absolute value of determinant, i.e. the point of highest dilatation, is $\boldsymbol{x} = A^{-1}\boldsymbol{c}$, $\boldsymbol{c} = (c_i)_{i=1,\ldots,d}^{\top}$.

3.2.3 Mean square differentiability

Let us now focus on differentiability questions. We give conditions for obtaining mean square differentiability (see section section 2.1.1). The random vector $\nabla Y_{\boldsymbol{x}} = \left(Y_i^{(1)}, \ldots, Y_d^{(1)}\right)^{\top}$, the gradient of a given Gaussian process Y at $\boldsymbol{x}$, will be used later in chapter 4 for the definition of new sampling criteria.

Proposition 7 (Mean-squared differentiability)**.** *Let Y be a centred Gaussian process with the covariance c defined in* (3.3)*. If*

- *for all $i \in \{1, \ldots, d'\}$, $\gamma_i(\cdot; \boldsymbol{\rho}_i) \in \mathcal{C}^1(\mathbb{R})$,*
- *for all $j, j' \in \{1, \ldots, d'\}$ and $\boldsymbol{u} \in \mathbb{R}^{d'}$, $\left.\frac{\partial^2 k_{\boldsymbol{b}}(\boldsymbol{v}, \boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{(\boldsymbol{u},\boldsymbol{u})}$ exists and is finite,*

then Y is mean-squared differentiable (i.e. has mean-squared derivatives in all canonical directions).

Proof. The tensor product T of the $\gamma_i(\cdot; \boldsymbol{\rho}_i)$ functions is of class $\mathcal{C}^1$ on $\mathbb{R}^d$. Using the regularity of $k_{\boldsymbol{b}}$ and T, and the chain rule applied to eq. (3.3) we obtain that $\forall \boldsymbol{x} \in \mathcal{D}$, $\forall i \in \{1, \ldots, d'\}$, $\left.\frac{\partial c(\boldsymbol{u}, \boldsymbol{u}')}{\partial u_i \partial u'_i}\right|_{(\boldsymbol{x},\boldsymbol{x})}$ exists. This property of c is equivalent to mean square differentiability (for centered GP, see e.g. Paciorek [2003] p. 49). □

3.2.4 Sample path differentiability

Another relevant property when defining a covariance function is the almost sure differentiability of sample paths of the associated GP.

Proposition 8 (Sample path differentiability)**.** *Consider the same assumptions as in proposition 7, and further assume that $\mathcal{D}$ is compact and there exist $C_0, \eta_0, \varepsilon_0 > 0$ such that $\forall j, j' \in \{1, \dots, d'\}$, and $\forall \boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}^{d'}$, $||\boldsymbol{u} - \boldsymbol{u}'|| < \varepsilon_0$, we have* $\left.\frac{\partial^2 k_{\boldsymbol{b}}(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{(\boldsymbol{u},\boldsymbol{u})} + \left.\frac{\partial^2 k_{\boldsymbol{b}}(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{(\boldsymbol{u}',\boldsymbol{u}')} - 2 \left.\frac{\partial^2 k_{\boldsymbol{b}}(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{(\boldsymbol{u},\boldsymbol{u}')} \leqslant \frac{C_0}{|\ln ||\boldsymbol{u}-\boldsymbol{u}'|||^{1+\eta_0}}$. *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 $0 < \varepsilon \leqslant 1/C_{\boldsymbol{\gamma}}$ (with $C_{\boldsymbol{\gamma}}$ a Lipschitz constant of $\boldsymbol{x} \to \boldsymbol{\gamma}(A\boldsymbol{x})$) such that:

1. $C = C_0 \sum_{j=1}^{q} \sum_{j'=1}^{q} a_{j1} a_{j'1} \sup_{\boldsymbol{x} \in \mathcal{D}} \left(\gamma'_j \left(\boldsymbol{a}_1^\top \boldsymbol{x}\right)\right) \sup_{\boldsymbol{x} \in \mathcal{D}} \left(\gamma'_{j'} \left(\boldsymbol{a}_1^\top \boldsymbol{x}'\right)\right)$,

2. $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$, $||\boldsymbol{x} - \boldsymbol{x}'|| < \varepsilon$ implies $||\boldsymbol{\gamma}(A\boldsymbol{x}) - \boldsymbol{\gamma}(A\boldsymbol{x}')|| < \varepsilon_0$ (by continuity of $\boldsymbol{\gamma}$),

3. $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$, $||\boldsymbol{x} - \boldsymbol{x}'|| < \varepsilon$ implies $\frac{1}{|\ln(C_{\boldsymbol{\gamma}} ||\boldsymbol{x}-\boldsymbol{x}'||)|^{1+\eta_0}} \leqslant \frac{1}{|\ln ||\boldsymbol{x}-\boldsymbol{x}'|||^{1+\eta}}$ (by existence of the limit $\lim_{h \to 0} \left(\frac{\ln |\ln |h||}{\ln |\ln(C_{\boldsymbol{\gamma}}) + \ln |h||}(1 + \eta_0) - 1\right) = \eta_0 > 0$).

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

$$
\begin{aligned}
&\left.\frac{\partial^2 c(\boldsymbol{u},\boldsymbol{u}')}{\partial u_1 \partial u'_1}\right|_{(\boldsymbol{x},\boldsymbol{x})} + \left.\frac{\partial^2 c(\boldsymbol{u},\boldsymbol{u}')}{\partial u_1 \partial u'_1}\right|_{(\boldsymbol{x}',\boldsymbol{x}')} - 2 \left.\frac{\partial^2 c(\boldsymbol{u},\boldsymbol{u}')}{\partial u_1 \partial u'_1}\right|_{(\boldsymbol{x},\boldsymbol{x}')} \\
&= \sum_{j=1}^{q} \sum_{j'=1}^{q} a_{j1} a_{j'1} \gamma'_j \left(\boldsymbol{a}_1^\top \boldsymbol{x}\right) \gamma'_{j'} \left(\boldsymbol{a}_1^\top \boldsymbol{x}'\right) \left(\left.\frac{\partial^2 k_{\boldsymbol{b}}(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{\substack{(\boldsymbol{\gamma}(A\boldsymbol{x}'), \\ \boldsymbol{\gamma}(A\boldsymbol{x}'))}} + \left.\frac{\partial^2 k_{\boldsymbol{b}}(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{\substack{(\boldsymbol{\gamma}(A\boldsymbol{x}), \\ \boldsymbol{\gamma}(A\boldsymbol{x}))}} - 2 \left.\frac{\partial^2 k_{\boldsymbol{b}}(\boldsymbol{v},\boldsymbol{v}')}{\partial v_j \partial v'_{j'}}\right|_{\substack{(\boldsymbol{\gamma}(A\boldsymbol{x}), \\ \boldsymbol{\gamma}(A\boldsymbol{x}'))}} \right) \\
&\leqslant \frac{C}{|\ln ||\boldsymbol{\gamma}(A\boldsymbol{x}) - \boldsymbol{\gamma}(A\boldsymbol{x}')|||^{1+\eta_0}} \leqslant \frac{C}{|\ln ||\boldsymbol{x} - \boldsymbol{x}'|||^{1+\eta}}. \qquad (3.20)
\end{aligned}
$$

Using the theorem of sample path continuity for GP derivatives (see e.g. Scheuerer [2009] p. 55), we get the sample path continuity for the GP $\partial Y / \partial x_1$ and thus ∇Y by generalising to all components. □

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

3.3 Examples of function approximation with WaMI-GP

In this section, we compare the performances of the WaMI-GP model in fixed design settings as well as in sequential design settings on an example function. The GP model by ordinary kriging as well as the maximum likelihood estimation (MLE) is computed in the R programming language with the package 'kergp' [Deville et al., 2015]. The gradient ascent method calculating the MLE is the L-BFGS-B algorithm implemented in R (function 'optim'). Note that, as we use the WaMI covariance within a standard form of GP modelling, the training cost has $O(n^3)$ complexity (due to an inversion of the matrix C in eq. (2.8)). TGP [Gramacy, 2007, Gramacy and Taddy, 2010] is an exception, where the division of the data set leads generally to a much faster parameter estimation. Significant efforts have been done to reduce computing efforts for standard GP models (see e.g. [Quiñonero-Candela and Rasmussen, 2005] and references therein), and this aspect was not a priority during our developments as we targeted applications with expensive-to-evaluate functions.

We first apply the WaMI-GP model to the synthetic data coming from the function in eq. (1.1) and we compare it with other models previously introduced: stationary anisotropic GP, axial warping GP, Treed GP. The prediction and posterior variance of the estimated models are shown in fig. 3.8. We observe that the estimated warping of the WaMI-GP dilated the input space in the high variation region, localised around the line of equation $2x_1+x_2 = 7/150$ (for the cliff).

We look now at the results of an MSE-driven sequential design of experiments under the WaMI-GP model (in standard setting) compared to three competitor models covered in the last chapter: stationary GP, GP with axial warping, and TGP. In fig. 3.9, sequential design are represented and in fig. 3.10 we display the absolute difference between the real function (example of eq. (2.48)) 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 (see the estimated warping in fig. 3.11) and also associating a higher MSE there, enables to comparatively achieve enhanced prediction performance as illustrated

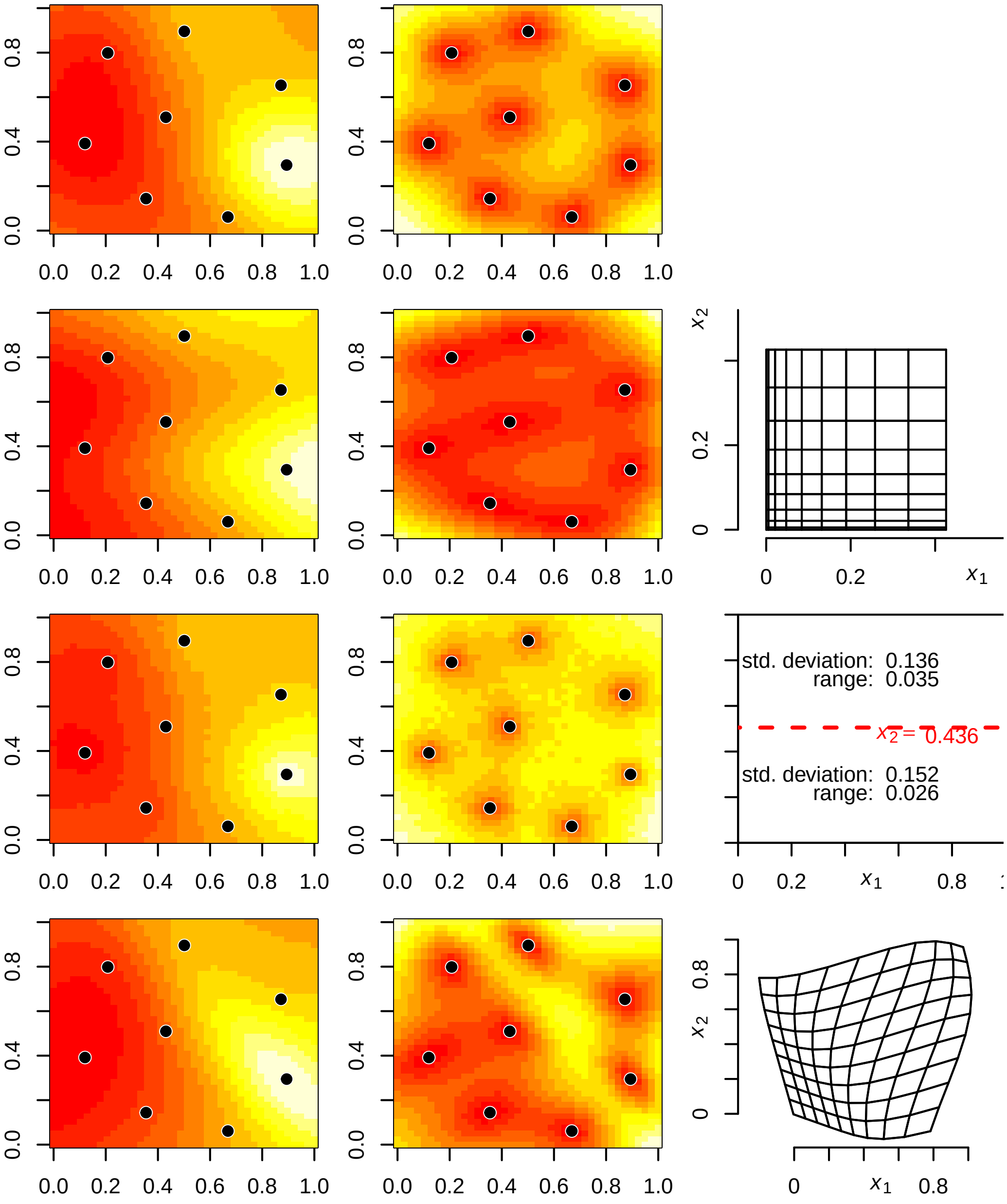


Figure 3.8: Application of different models on a synthetic data set (eq. (1.1)). The models are stationary anisotropic, axial warping, TGP and WaMI-GP. The predictions are represented in the first column of images. The second column is for the posterior standard deviation. The third column represents features of the models. For axial warping and WaMI-GP models, the overall warping is represented using the warping of a 9 × 9 regular grid. For the TGP model, we show the partition of the space, the standard deviations and ranges.

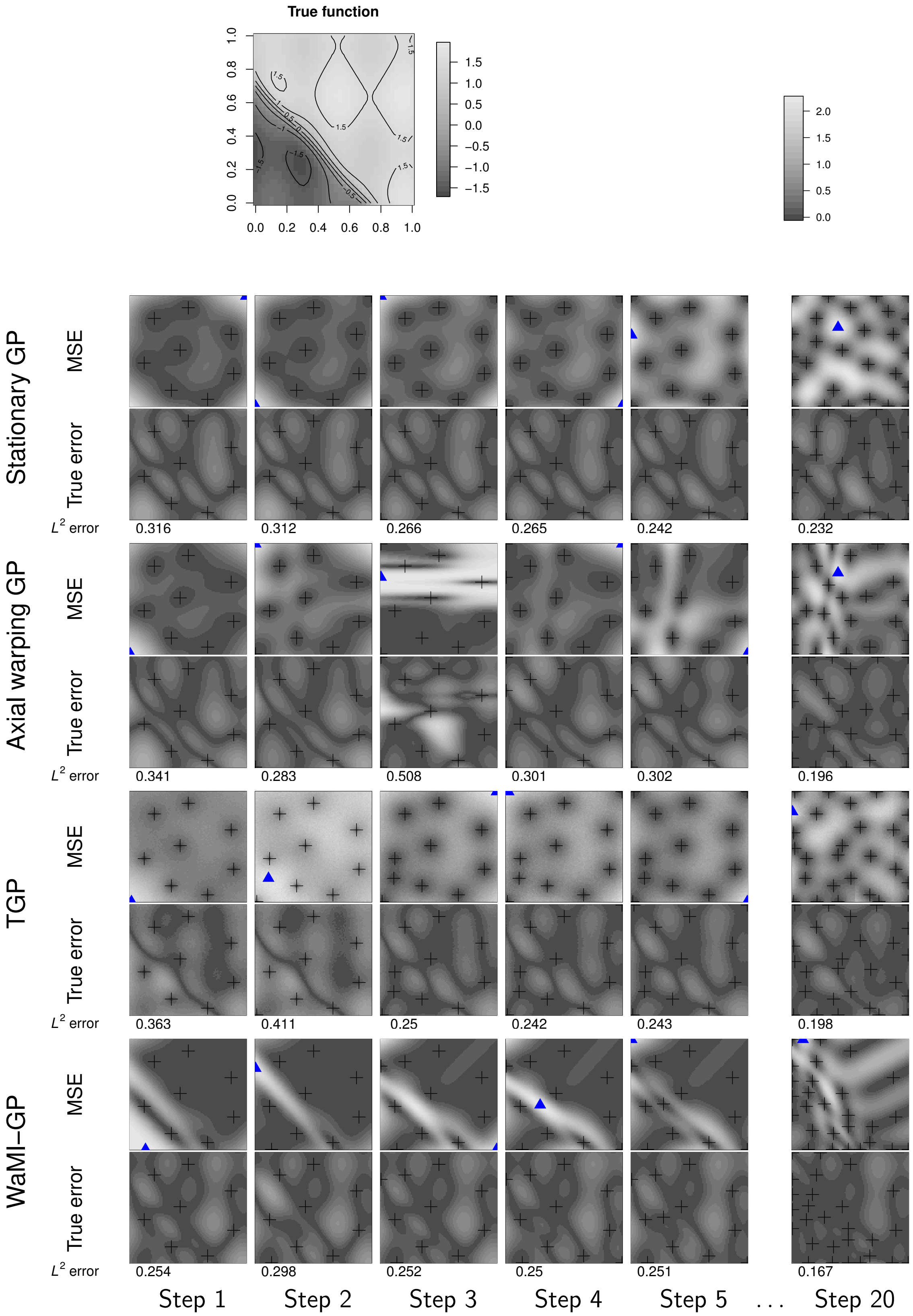


Figure 3.9: Model-based sequential design of experiments of a function $f : \boldsymbol{x} \in [0,1]^2 \to \frac{\sin(15x_1)+\cos(10x_2)}{5} + \arctan\left(\frac{20(x_1+x_2)-15}{2}\right)$. The different models are stationary anisotropic, axial warping, TGP and WaMI-GP. For each method, we see the first 5 steps and last step. The point for the next evaluation (blue triangle) is MSE optimal.

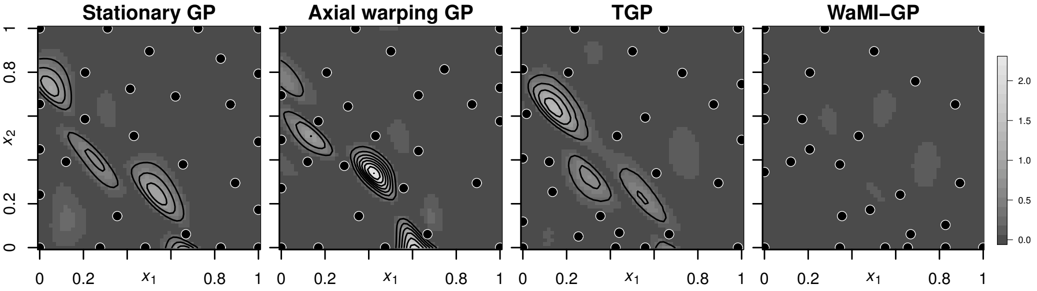


Figure 3.10: Prediction errors of four competing models on the running example function. The different models are a stationary anisotropic GP, an axial warping GP, Treed GP and WaMI-GP. For each method, we see the tenth step of a sequential design driven by the MSE criterion (shared initial design).

in fig. 3.12.

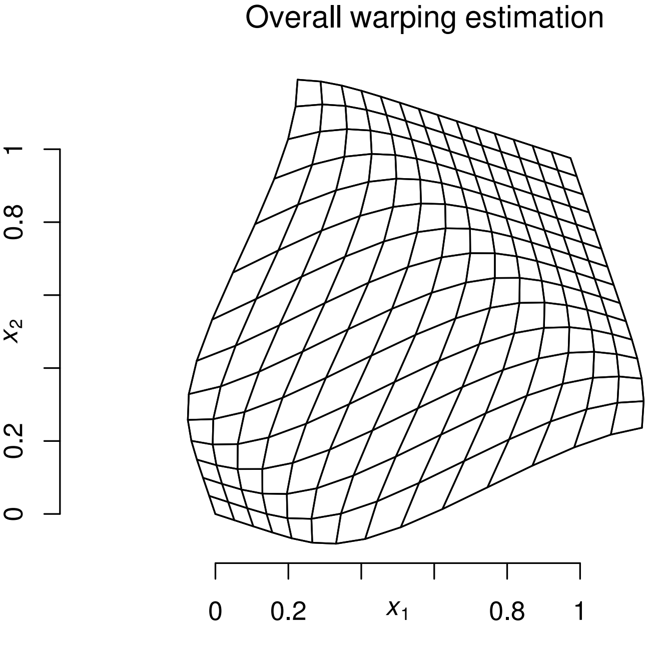


Figure 3.11: Estimated warping from the WaMI-GP model at step 20 of a MSE driven design of experiments.

These numerical results illustrate that WaMI-GP is able to account for heterogeneous regions in a semi-automated way. Here all parameters including axes are estimated by maximum likelihood but the initial kernel k_β and the number of warping dimensions is fixed in advance. Hence WaMI-GP improves

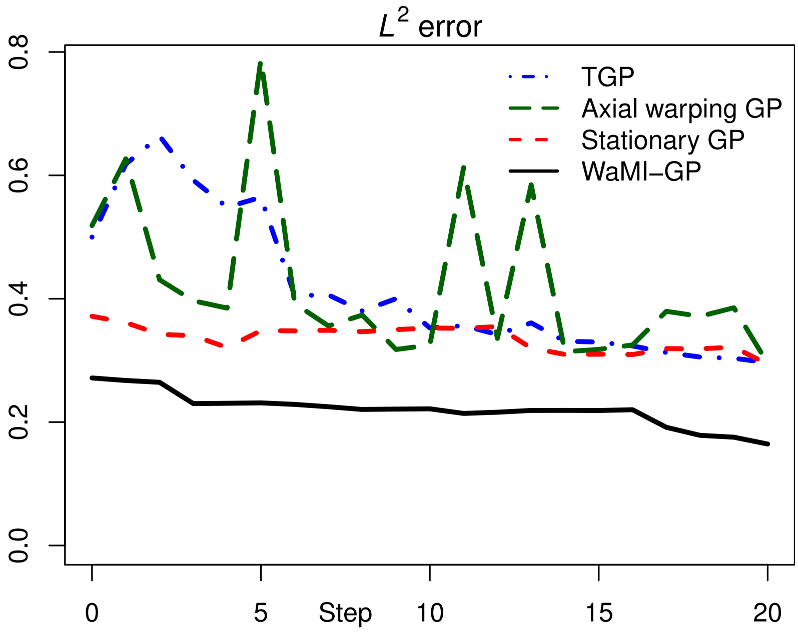


Figure 3.12: Prediction error of the running example function by the four considered models at each step of MSE-driven sequential designs.

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 chapter 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 any GP model. Later on in chapter 5 the two approaches of working on kernels and/or on the sampling criteria for learning functions with heterogeneous variations will be combined and compared in a benchmark study.

3.4 Non-parametric warping estimation using local scale analysis

The objective is to exploit the local scale (section 2.2.4) provided by a wavelet transform of a GP model to approach an input space warping. In a typical *Computer Experiments* framework, one of the difficulties for computing (2.51) stands in having a scarce design which is generally not a regular grid. To circumvent this limitation, we propose an original warping approximation algorithm that is tractable when considering scattered data. The starting point is local scale estimation (section 2.2.4) relying on a Gaussian process Y conditioned on the available evaluation results, represented by the event $\mathcal{A}_n = \{Y_{x_1} = f(x_1), \ldots, Y_{x_n} = f(x_n)\}$. A clear advantage of this approach is that it allows local estimation without any constraint on the evaluation design while also benefiting from the versatility and the tractability of GP models Roustant et al. [2012]. Here the probabilistic nature of the model is exploited when performing local scale estimation, as the expectations coming into play in section 2.2.4 are estimated by Monte Carlo simulations relying on the posterior GP distribution. Our proposed approach consists of stationarising the GP model by successive index transformations relying on this local scale estimation scheme. The proposed approach is summarised by the Algorithm 3.4 below:

Algorithm

Inputs: fixed number of steps N_{stop}, number of Monte Carlo samples p, a prior distribution for the GP Y (stationary, by default) and the current set of evaluation results $\mathcal{A}_n$.

Start.

```
– Build a GP model by conditioning $Y$ to the data $\mathcal{A}_n$ using the kriging equations. Set $Y^{(0)} = Y$.
– For $i = 0, \ldots, N_{\text{stop}}$:
    • Sample $p$ realisations $\{y_n^{(i)}\}_{n=1,\ldots,p}$ of $Y^{(i)}$ conditional on $\mathcal{A}_n$,
    • Apply the wavelet transform to evaluate $\{\mathcal{W}_{y_n^{(i)}}(\tau, s)\}_{n=1,\ldots,p}$
      and estimate $\mathbb{E}\left(|\mathcal{W}_{Y^{(i)}}(\tau, s)|^2\right)$,
    • Compute the local scale (section 2.2.4) to get $\left(\gamma^{(i)}\right)'(\tau) = G_{Y^{(i)}}(\tau)$ and
      thus $\gamma^{(i)}(\tau)$ by numerical integration,
    • 'Stationarise' $Y^{(i)}$, i.e. consider a new process $Y^{(i+1)} = D_{\gamma^{(i)-1}} Y^i$
      obtained by warping $Y^{(i)}$ using $\gamma^{(i)-1}$.
```

End.

The estimated overall warping function γ is the chaining of the warpings computed step by step,

$$\gamma = \gamma^{(N_{\text{stop}})} \circ \ldots \circ \gamma^{(0)}. \tag{3.21}$$

In an attempt to give an interpretation of this algorithm, let us consider its first iteration. It leads to a warping function satisfying, for any $\tau \in \mathbb{R}$, $\left(\gamma^{(0)}\right)'(\tau) = G_{Y^{(0)}}(\tau)$. Let us assume that the inverse warping operator in the vicinity of τ can be replaced by its tangent at τ due to narrow localisation of the wavelet. Neglecting the errors associated to the estimation of the mathematical expectation and of the local scale and replacing the slope coefficient a in section 2.2.4 by the slope of the tangent of $\gamma^{(0)-1}$ at τ delivers

$$\left(\gamma^{(1)}\right)'(\tau) = G_{D_{\left(\gamma^{(0)}\right)^{-1}} Y^0}(\tau) \approx \left(\gamma^{(0)-1}\right)'(\tau) \times \gamma^{(0)\prime}\left(\gamma^{(0)-1}(\tau)\right) = 1.$$

This idealised settings yields convergence of the algorithm after only one iteration (we have $\tau \in \mathbb{R}$, $i \geqslant 1$, $(\gamma^{(i)})'(\tau) = 1$). As we neglect error terms in the approximation of $\gamma^{(i)-1}(x)$, the algorithm does not actually converge in one iteration. The last equation is rather to be considered as a colloquial explanation of the numerical observation that local scale estimates get closer to a constant over iterations, corresponding to a stepwise stationarisation of the GP. Numerical investigations about the behaviour of the algorithm are presented in section 5.3.

Chapter 4

Sampling criteria for adaptive designs of experiments

We describe in this chapter two contributions in adaptive sampling, related to prediction of functions with heterogeneous variations (section 4.1) and to global optimisation (section 4.2). For each of these issues, we use the derivatives of a GP conditioned on evaluations. First, as detailed earlier in section 4.1, the GP derivatives are a key tool for the detection of high-variation zones. Second, GP derivatives are involved in the calculation of the gradient of a sampling criteria, the multipoint expected improvement (see appendix A for a detailed introduction on the multipoint EI criterion). We first recall that under sufficient regularity conditions, the gradient $\nabla Y \triangleq (\nabla Y_{\boldsymbol{x}})_{\boldsymbol{x} \in \mathcal{D}}$ is a vector-valued Gaussian Process. The conditional distribution knowing n evaluations $\mathcal{A}_n$ is driven by derivatives of m_n and c_n (see the properties on GP differentiation introduced in section 2.1.1). In the following section, the distribution of $\nabla Y_{\boldsymbol{x}}$ is used for building sampling criteria dedicated to the exploration of functions with heterogeneous variations.

4.1 Learning heterogeneous functions

4.1.1 Core idea

We look for sampling criteria answering the following problem: how to distribute a limited evaluation budget in order to reduce the prediction error of a function with heterogeneous variations. In particular, we are interested in algorithms which would automatically detect regions with higher variations in order to allocate a larger proportion of budget to these regions.

In chapters 2 and 3, we approximated some functions by diverse GP-related methods and concluded that exploring high-variation regions was key to quickly reducing 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 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 show more local variability. A problem however with variance-based criteria such as considered so far is that they are homoscedastic in the observations, or in other words, they depend solely on the geometry of the experimental design and not on the response values. Hence trying to locate high-variation regions with variance-based criteria does not make much sense, unless the model accounts for heterogeneities through estimated parameters that reflect them, such as with WaMI-GP. Our approach here is to rely instead on the gradient of the GP in order to add points in unexplored regions with potentially high slopes.

Different scalar indicators quantifying local variations and related uncertainties could be defined. We chose to focus essentially on variance-based criteria for (exponentiated) gradient norms. Toward this means, let us consider the squared gradient norm process $(Q_{\boldsymbol{x}})_{\boldsymbol{x}\in\mathcal{D}}$ defined by

$$Q_{\boldsymbol{x}} = ||\nabla Y_{\boldsymbol{x}}||^2_{\mathbb{R}^d} = \nabla Y_{\boldsymbol{x}}^\top \nabla Y_{\boldsymbol{x}}. \tag{4.1}$$

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. This problem involves the probability distribution of quadratic forms in arbitrary Gaussian variables. Yet, as we develop next, some (fractional) moments of $Q_{\boldsymbol{x}}$ can be calculated in closed form or computed efficiently, leading to practical infill sampling criteria.

4.1.2 Definitions and calculations of gradient-based infill criteria

We propose several infill criteria based on the gradient norm process and provide calculus elements for their fast computation. Let us consider the MSE and IMSE criteria applied to the exponentiated gradient norm.

Definition 6 (Gradient Norm Variance criterion and generalisations)**.** *Given n function evaluation results $\mathcal{A}_n$ and $\boldsymbol{x} \in \mathcal{D}$, we define the Gradient Norm Variance (GNV) criterion as*

$$J_n^{\mathrm{GNV}}(\boldsymbol{x}) = \mathrm{var}\left(||\nabla Y_{\boldsymbol{x}}|| \mid \mathcal{A}_n\right) = \mathrm{var}\left(\sqrt{Q_{\boldsymbol{x}}} \,\middle|\, \mathcal{A}_n\right) \tag{4.2}$$

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

$$J_n^{\mathrm{GNV},\eta}(\boldsymbol{x}) = \mathrm{var}\left(||\nabla Y_{\boldsymbol{x}}||^{\eta} \mid \mathcal{A}_n\right) = \mathrm{var}\left(Q_{\boldsymbol{x}}^{\eta/2} \mid \mathcal{A}_n\right). \tag{4.3}$$

This class of criteria can also be generalized in the same way as IMSE *generalizes* MSE, *by integration. Indeed, as mentioned in section 2.1.2, the* IMSE *value at $\boldsymbol{x} \in \mathcal{D}$ is the integral of the expected variance if the next evaluation is $\boldsymbol{x}$. Similarly, we define the* IGNV *criterion by*

$$J_n^{\mathrm{IGNV},\eta}(\boldsymbol{x}) = \int\limits_{\boldsymbol{u}\in\mathcal{D}} \mathbb{E}\left(\mathrm{var}\left(Q_{\boldsymbol{u}}^{\eta/2} \mid \mathcal{A}_n, Y_{\boldsymbol{x}}\right) \middle| \mathcal{A}_n\right) \mathbf{d}\boldsymbol{u}. \tag{4.4}$$

While the transformed norm loses its homogeneity because of the exponent, we still refer to this criterion as a norm variance: “GNV with exponent η” or “GNV(η)”. GNV(1) is hence the previous GNV, and we will also pay a particular attention to GNV(2) in what follows.

The following property gives a closed form 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 9. *Let $\boldsymbol{x} \in \mathcal{D}$ and denote by $(\lambda_i(\boldsymbol{x}))_{1\leqslant i\leqslant d}$ the eigenvalues of $\nabla\otimes\nabla^{\top} c_n(\boldsymbol{x},\boldsymbol{x})$. Then, the* GNV(2) *criterion can be written as follows:*

$$J_n^{\mathrm{GNV},\eta=2}(\boldsymbol{x}) = 4\ \nabla m_n(\boldsymbol{x})^{\top}\nabla\otimes\nabla^{\top} c_n(\boldsymbol{x},\boldsymbol{x})\nabla m_n(\boldsymbol{x}) + 2\sum_{i=1}^{d}\lambda_i(\boldsymbol{x})^2. \tag{4.5}$$

Furthermore, the GNV(1) *criterion can be expanded as follows:*

$$J_n^{\mathrm{GNV},\eta=1}(\boldsymbol{x}) = ||\nabla m_n(\boldsymbol{x})||^2 + \mathrm{tr}\left(\nabla\otimes\nabla^{\top} c_n(\boldsymbol{x},\boldsymbol{x})\right) - \mathbb{E}\left(\sqrt{Q_{\boldsymbol{x}}}|\mathcal{A}_n\right)^2. \tag{4.6}$$

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

$$J_n^{\mathrm{IGNV},\eta=1}(\boldsymbol{x}) = \int\limits_{D} \left(||\nabla m_n(\boldsymbol{u})||^2 + \frac{1}{c_n(\boldsymbol{x},\boldsymbol{x})} \kappa_n(\boldsymbol{u},\boldsymbol{x})^\top \kappa_n(\boldsymbol{u},\boldsymbol{x}) \right) \mathbf{d}\boldsymbol{u}$$
$$+ \int\limits_{D} \left(\mathrm{tr}\left(\nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \right) - \mathbb{E}\left(\mathbb{E}\left(\sqrt{Q_{\boldsymbol{u}}} | \mathcal{A}_n, Y_{\boldsymbol{x}} \right)^2 \middle| \mathcal{A}_n \right) \right) \mathbf{d}\boldsymbol{u}. \tag{4.7}$$

In the case $\eta = 2$*, we have*

$$J_n^{\mathrm{IGNV},\eta=2}(\boldsymbol{x}) = \int\limits_{\boldsymbol{u}\in\mathcal{D}} \left(4 \nabla m_n(\boldsymbol{u})^\top \nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \nabla m_n(\boldsymbol{u}) + 2 \sum_{i=1}^{d} \lambda_{i,\boldsymbol{x}}(\boldsymbol{u})^2 \right) \mathbf{d}\boldsymbol{u}$$
$$+ \frac{4}{\mathrm{var}\left(Y_{\boldsymbol{x}} | \, \mathcal{A}_n \right)} \int\limits_{\boldsymbol{u}\in\mathcal{D}} \kappa_n(\boldsymbol{u},\boldsymbol{x})^\top \nabla \otimes \nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \kappa_n(\boldsymbol{u},\boldsymbol{x}) \, \mathbf{d}\boldsymbol{u} \tag{4.8}$$

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* $Y_{\boldsymbol{x}}$ *given* $\mathcal{A}_n$*.*

Proof of proposition 9. Let us first address the case $\eta = 1$ using the notation $\boldsymbol{Z}^c_{\boldsymbol{x}} = \boldsymbol{Z}_{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}}$ with $\boldsymbol{m}_{\boldsymbol{x}} = \nabla m_n(\boldsymbol{x})$. The first step is to expand the criterion as follows:

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

The term $\mathrm{var}\left(\boldsymbol{Z}^{c\,\top}_{\boldsymbol{x}} \boldsymbol{Z}^c_{\boldsymbol{x}} \right)$ can be further expanded as $\boldsymbol{Z}^c_{\boldsymbol{x}} = U_{\boldsymbol{x}} D_{\boldsymbol{x}}^{\frac{1}{2}} \boldsymbol{N}$ 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:

$$\begin{aligned}\mathrm{var}\left(||\boldsymbol{Z}_{\boldsymbol{x}}||\right) &= \mathbb{E}\left(\boldsymbol{Z}_{\boldsymbol{x}}^\top \boldsymbol{Z}_{\boldsymbol{x}}\right) - \mathbb{E}\left(\sqrt{\boldsymbol{Z}_{\boldsymbol{x}}^\top \boldsymbol{Z}_{\boldsymbol{x}}}\right)^2 \\ &= \boldsymbol{m}_{\boldsymbol{x}}^\top \boldsymbol{m}_{\boldsymbol{x}} + \underbrace{2\boldsymbol{m}_{\boldsymbol{x}}^\top \mathbb{E}\left(\boldsymbol{Z}_{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}}\right)}_{=0} + \sum_{i=1}^{d} \mathrm{var}\left([\boldsymbol{Z}_{\boldsymbol{x}}]_i\right) - \mathbb{E}\left(\sqrt{Q_{\boldsymbol{x}}}\right)^2 .\end{aligned} \tag{4.9}$$

For the proof of $\mathrm{IGNV}_{\eta=1,2}$, we focus on the integrand. We formulate the case $\eta = 2$ by applying on eq. (4.5) the operation "$\mathbb{E}\left(\cdot\middle|\, \mathcal{A}_n\right)$", i.e.

$$\begin{aligned}&\mathbb{E}\left(\mathrm{var}\left(Q_{\boldsymbol{u}} \mid \mathcal{A}_n, Y_{\boldsymbol{x}}\right)\middle| \mathcal{A}_n\right) = \qquad (4.10)\\ &4\mathbb{E}\left(\mathbb{E}\left(\nabla Y_{\boldsymbol{u}}\middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right)^\top \nabla\otimes\nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u}) \mathbb{E}\left(\nabla Y_{\boldsymbol{u}}\middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right)\middle| \mathcal{A}_n\right) + 2\sum \lambda_{i,\boldsymbol{x}}(\boldsymbol{u})^2.\end{aligned}$$

The second term does not get affected as the covariance matrix $\nabla\otimes\nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u})$ is deterministic. Then we get the result with the following formula derived from a Gaussian vector conditioning

$$\mathbb{E}\left(\nabla Y_{\boldsymbol{u}}\middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right) = \nabla \boldsymbol{m}_n(\boldsymbol{u}) + \frac{Y_{\boldsymbol{x}} - m_n(\boldsymbol{x})}{c_n\left(\boldsymbol{x},\boldsymbol{x}\right)} \kappa_n(\boldsymbol{u},\boldsymbol{x}). \tag{4.11}$$

For $\eta = 1$, using the result of eq. (4.9), we obtain

$$\begin{aligned}\mathbb{E}\left(\mathrm{var}\left(\sqrt{Q_{\boldsymbol{u}}}\ \middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right)\middle| \mathcal{A}_n\right) &= \mathbb{E}\left(||\mathbb{E}\left(\nabla Y_{\boldsymbol{u}}\middle| \mathcal{A}_n, Y_{\boldsymbol{x}}\right)||^2\middle| \mathcal{A}_n\right) \\ &+ \mathrm{tr}\left(\nabla\otimes\nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u})\right) - \mathbb{E}\left(\mathbb{E}\left(\sqrt{Q_{\boldsymbol{u}}}|\mathcal{A}_n, Y_{\boldsymbol{x}}\right)^2\middle| \mathcal{A}_n\right).\end{aligned} \tag{4.12}$$

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

□

The expectation terms in eqs. (4.6) and (4.7) can be approximated by quadrature formulas of univariate or bivariate integrals: for $\boldsymbol{u}$, $\boldsymbol{x}$ in $\mathcal{D}$,

$$\mathbb{E}\left(\sqrt{Q_{\boldsymbol{x}}}\right) = \int_{\mathbb{R}} \sqrt{t} f_Q\left(t; \nabla m_n(\boldsymbol{x}), \nabla\otimes\nabla^\top c_n(\boldsymbol{x},\boldsymbol{x})\right) \mathrm{d}t \tag{4.13}$$

$$\begin{aligned}&\mathbb{E}\left(\mathbb{E}\left(\sqrt{Q_{\boldsymbol{u}}}|\mathcal{A}_n, Y_{\boldsymbol{x}}\right)^2\middle| \mathcal{A}_n\right) = \\ &\int_{\mathbb{R}} \left(\int_{\mathbb{R}} \sqrt{t} f_Q\left(t; \boldsymbol{\mu}_n(y;\boldsymbol{u},\boldsymbol{x}), \Gamma_n(\boldsymbol{u},\boldsymbol{x})\right) \mathrm{d}t\right)^2 \varphi_{1,c_n(\boldsymbol{x},\boldsymbol{x})}(y - m_n(\boldsymbol{x}))\mathrm{d}y, \qquad (4.14)\end{aligned}$$

where $\boldsymbol{\mu}_n(y;\boldsymbol{u},\boldsymbol{x}) = \nabla m_n(\boldsymbol{u}) + \frac{y-m_n(\boldsymbol{x})}{c_n(\boldsymbol{x},\boldsymbol{x})}\kappa_n(\boldsymbol{u},\boldsymbol{x})$, $\Gamma_n(\boldsymbol{u},\boldsymbol{x}) = \nabla\otimes\nabla^\top c_{n,\boldsymbol{x}}(\boldsymbol{u},\boldsymbol{u})$ and $\varphi_{1,c_n(\boldsymbol{x},\boldsymbol{x})}(\cdot - m_n(\boldsymbol{x}))$ is 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 [Duchesne and De Micheaux, 2010] and implemented in a R package *CompQuadForm* [Duchesne and De Micheaux, 2010]. The most recent method is based on approximating with a distribution of a *central* quadratic form, tuned for egalizing the three first moments [Pearson, 1959] (equal skewness). The method of [Liu et al., 2009] is closely related to this, and provides error bounds. There also exist methods for which the error can be made arbitrarily small, that use numerical inversion of the characteristic function [Imhof, 1961] or an infinite series formulation [Farebrother, 1984].

4.1.3 Observations on examples

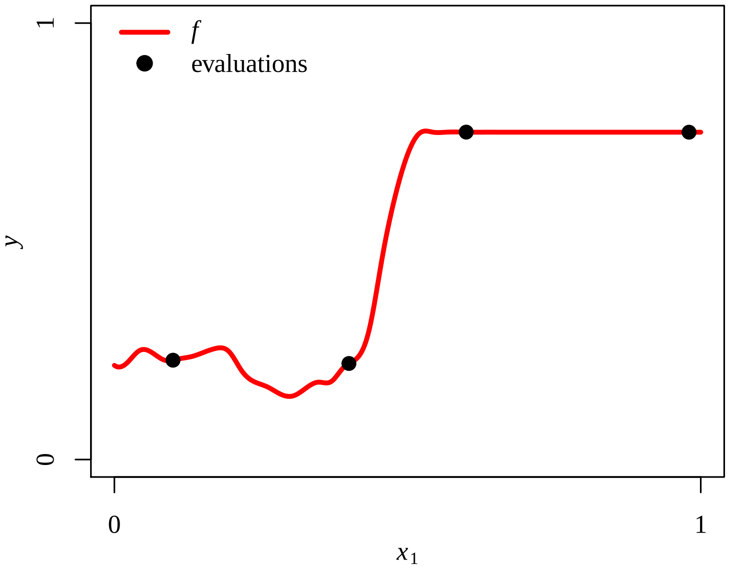


Figure 4.1: Objective function and evaluations for the univariate example.

We compute and display the proposed criteria for two synthetic test cases in dimension one and two.

The univariate test case is a first simple illustration with $\mathcal{D} = [0,1]$. We construct two models from four arbitrary evaluation points $X_{1:4} = (0.1, 0.4, 0.6, 0.98)^\top$. Evaluation values $\boldsymbol{y}_{1:4} = (0.225, 0.22, 0.7501, 0.75)^\top$ are taken from a IRSN test case, with a steep slope in the middle of $\mathcal{D}$ (see fig. 4.1, with contextual details

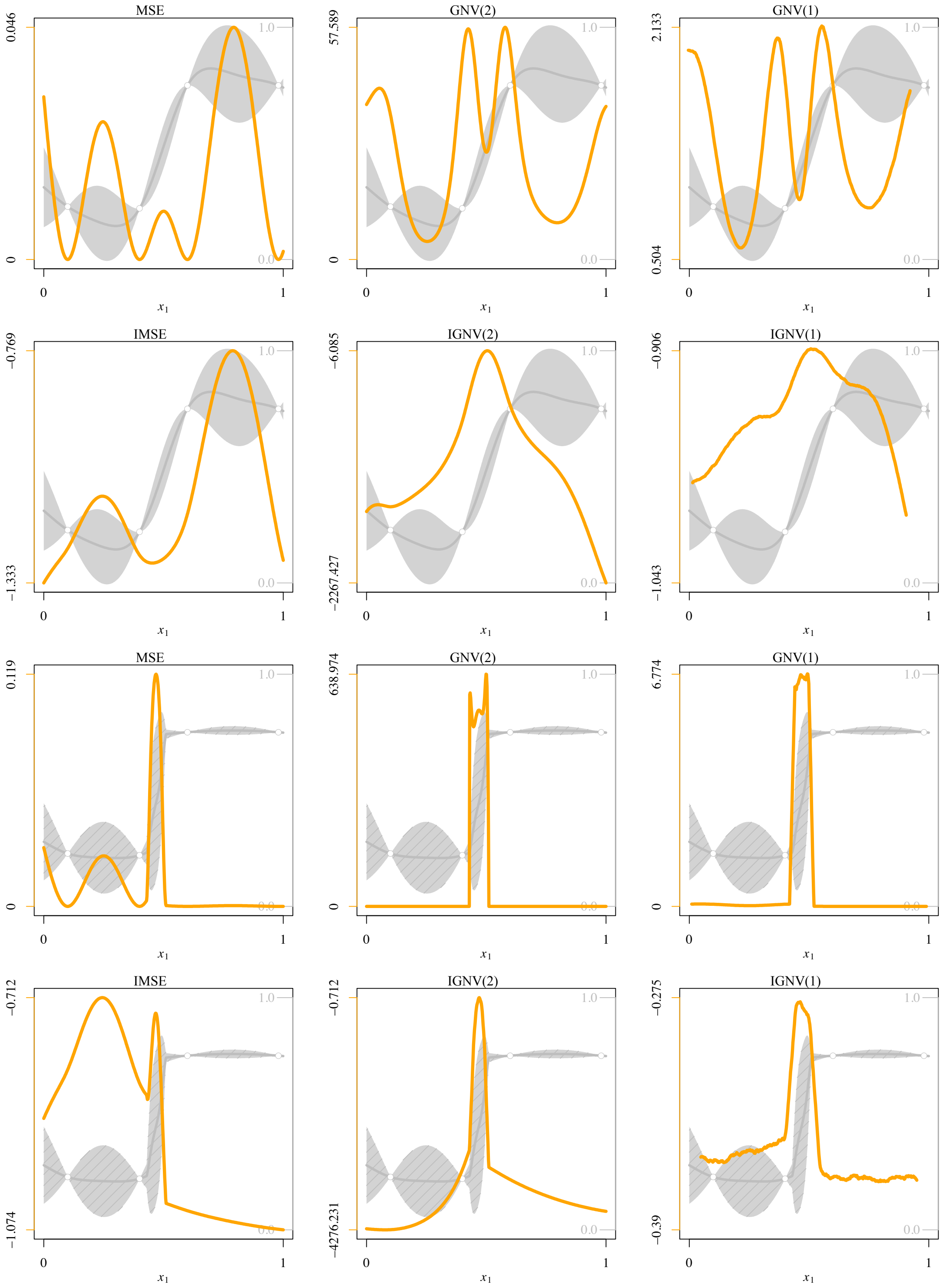


Figure 4.2: Criteria values (orange) on $\mathcal{D}$. The criteria are applied on two Gaussian process models, with mean and standard deviation represented in the background. Plain grey shows stationary model and crosshatched grey shows non-stationary.

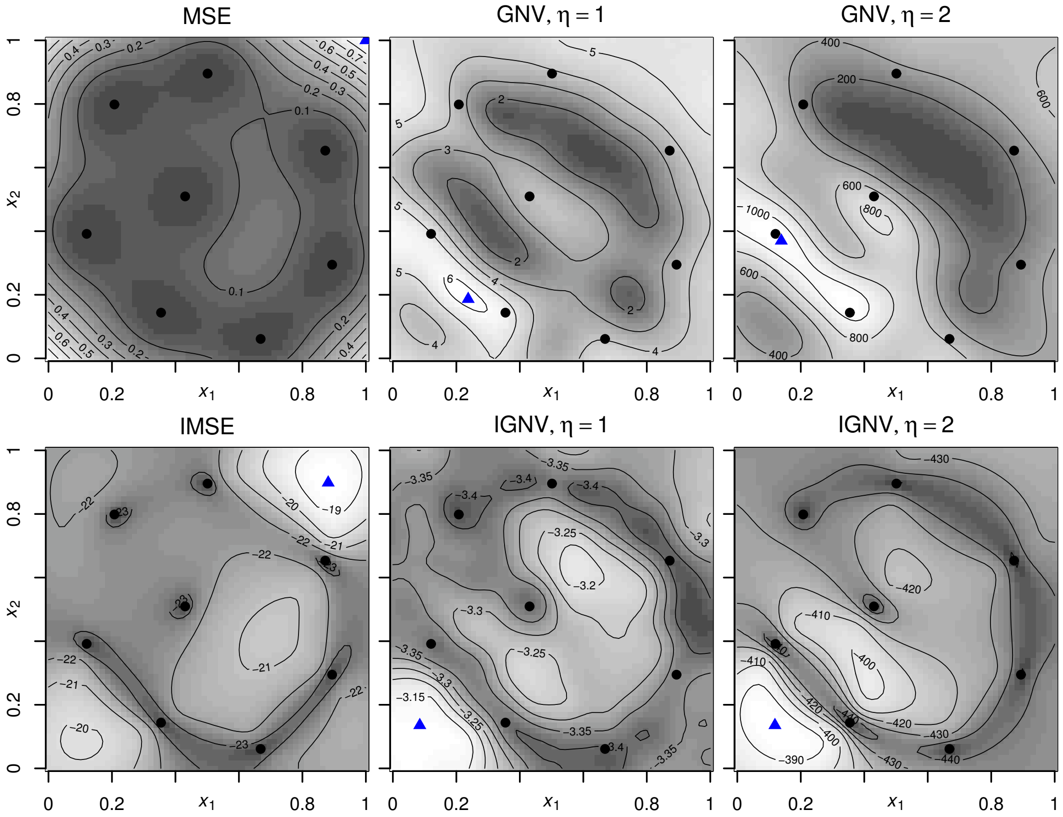


Figure 4.3: Classical and proposed criteria according to a stationary Gaussian process modelling.

in section 5.1). As the criteria values depend on the model, and in particular on the covariance function, we build two models for this data set via ordinary kriging. The first model has a stationary covariance, with Matérn structure (smoothness parameter $\nu = 5/2$ and correlation length $\theta = 2/10$). The second model is identical, but with a non-stationary covariance obtained by chaining the the first covariance with a warping γ as in section 2.2.2, with γ defined as follows:

$$\gamma : x \to x\mathbb{1}_{[0,4.5[}(x)+(10x-4.5+0.45)\mathbb{1}_{[4.5,5.5[}(x)+(0.1(x-0.55)+1.45)\mathbb{1}_{[5.5,1]}(x). \tag{4.15}$$

This warping correspond to a contraction of $[0, 1]$ in its right side and a dilatation in its middle. Figure 4.2 displays the values of six criteria, MSE, IMSE, GNV and IGNV for $\eta = 1, 2$ for the stationary and the non-stationary models.

The bivariate example approximates the test function of eq. (2.48) displayed in fig. 2.9 with the GP model stationary isotropic of Matérn covariance (section 2.1) from a LHS design of size 8 optimised with the maximin distance. Figure 4.3 display the values of the MSE, IMSE GNV and IGNV for $\eta = 1, 2$.

We observe for both uni- and bivariate cases that despite their higher computational costs, integrated criteria (IMSE, IGNV) 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, variance-based criteria do not provide a higher criterion value for the high variation region in the bottom left corner of the input space. On the contrary, we notice that gradient-based criteria provide higher values where f has high variations.

Like MSE and IMSE, integrated gradient-based criteria insure that new evaluation points are not in regions surrounding points already evaluated, as they take the least optimal values in these region. These figures suggest that integrated gradient-based criteria could be useful as a compromise between global uncertainty reduction and focus on high variations. The univariate case also shows the impact of the model on the criterion value. Here the difference between the models is an input space warping. We observe, for the non-stationary model, that the contraction in the right sides of $[0, 1]$ (respectively the dilatation in the middle) lowers (respectively raises) the criterion value. These figures suggest that fitting a non-stationary warped model to a function with high variation regions will also make the criteria evaluate more in these regions, especially for the gradient-based criteria. These points will be investigated in the next chapter, where the different approaches developed throughout the thesis are tested and compared based engineering test cases.

4.2 Contributions in parallel Bayesian optimisation

We focus now on the multipoint expected improvement criterion (see appendix A for a detailed introduction). Our aim is to present a set of novel analytical and numerical results related to the calculation, the computation, and the maximisation of the multipoint EI criterion. As most of these results apply to a broader class of criteria, we present first in section 4.2.1 a generalisation of the multipoint EI that allows accounting for noise in conditioning observations and also exponentiating the improvement. This generalised criterion is calculated using moments of truncated Gaussian vectors in the flavor of [Chevalier and Ginsbourger, 2014.]. Thus in section 4.2.1 we give a calculus method for the derivation of these moments at any order. After this preliminary result, the generalised multipoint EI is calculated in section 4.2.1. The obtained formula is then revisited in the standard case (noise-free with an exponent set to 1), leading in section 4.2.2 to a numerical approximation of the multipoint EI with arbitrary precision and very significantly reduced computation time. Next, the (qd)-dimensional maximisation of the multipoint EI criterion is discussed in section 4.2.2, where the differentiability of the generalised criterion is studied and its analytical gradient is calculated. A numerical approach for fast gradient approximations with controllable accuracy is presented in section 4.2.2. Finally, section 4.2.2 discusses the computational gain of the proposed approaches.

4.2.1 Analytic results for generalised multipoint expected improvement

Generalisation of the multipoint expected improvement criterion

Throughout this section the objective function f may be observed noise-free or with noise, meaning that at some arbitrary iteration i the observed value may be $f(\boldsymbol{x}_i)$ or $f(\boldsymbol{x}_i) + \varepsilon_i$ where ε_i is a realisation of a zero mean Gaussian random variable with known (or estimated and plugged-in, see explanations on empirical Bayes approach in section 2.1.1) variance. We recall that f is assumed to be one realisation of a GP Y conditionally to evaluation events $\mathcal{A}_n \triangleq \{Y_{\boldsymbol{x}_1} = f(\boldsymbol{x}_1), \ldots, Y_{\boldsymbol{x}_n} = f(\boldsymbol{x}_1)\}$ (with conditioning on $Y_{\boldsymbol{x}_i} + \varepsilon_i$ in the noisy case). In noisy cases the ε_i's are generally assumed to be independent (although the case of ε_i's forming a Gaussian vector is tractable), but more essentially they are assumed independent of Y.

In batch-sequential Bayesian methods, in particular for optimisation, we are interested in computing sampling criteria J_n depending on $q \geqslant 1$ new points $\boldsymbol{x}_{n+1:n+q} = (\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q}) \in \mathcal{D}^q$. At any step of corresponding (synchronous) parallel algorithms, the next batch of q points $\boldsymbol{x}^{\star}_{n+1:n+q}$ is then defined by maximizing J_n over all possible batches.

Values of such criteria typically depend on $\boldsymbol{x}_{n+1:n+q}$ through the conditional distribution of $Y_{\boldsymbol{x}_{1:n+q}}$ knowing $\mathcal{A}_n$, which in noiseless setting simplifies to $Y_{\boldsymbol{x}_{n+1:n+q}}$ (as without noise, $Y_{\boldsymbol{x}_{1:n}}$ is deterministic knowing $\mathcal{A}_n$). Conditional mean and covariance functions are analytically formulated via the Kriging equations, see section 2.1.1. Working out these criteria thus generally boils down to Gaussian vector calculus, which may become intricate and quite cumbersome to implement as q (or $n + q$, in noisy settings) increases. Our generalised version of the multipoint expected improvement criterion (or 'q-EI', when the evaluation batch has size q), that allows accounting for a Gaussian noise in the conditioning observations and also for an exponentiation in the definition of the improvement, is defined as:

$$\mathrm{EI}_n(\boldsymbol{x}_{n+1:n+q}) = \mathbb{E}_n\left(\left(\min_{\ell=1,\ldots,n} Y_{\boldsymbol{x}_\ell} - \min_{r=1,\ldots,q} Y_{\boldsymbol{x}_{n+r}}\right)_+^{\alpha}\right), \tag{4.16}$$

where $\alpha \in \mathbb{N}^{\star}$, $\mathbb{E}_n(\cdot) = \mathbb{E}(\cdot|\mathcal{A}_n)$ and $(\cdot)_+ \triangleq \max(0,\cdot)$. This form gathers several sampling criteria notably including q-EI, both in noiseless and noisy settings, and also a multipoint version of the generalised EI of [Schonlau, 1997]. In addition, the obtained results apply to batch-sequential versions of the Expected Quantile Improvement [Picheny et al., 2013] (EQI) and variations thereof, by a simply change of process from Y to the quantile process. We will show in proposition 11 that such generalised multipoint EI criteria can be formulated as a sum of moments of truncated Gaussian vectors. In the next subsection, in order to get a closed form for the generalised EI we first define these moments and derive some first analytical formulas, that might also be of relevance in further contexts.

Preliminary calculations on moments of truncated Gaussian distribution

We fix $\alpha \in \mathbb{N}^{\star}$ and $p = n + q$ in noisy settings or $p = q$ in noiseless settings.

Definition 7. *Let $\boldsymbol{Z}$ be a Gaussian vector with mean $\boldsymbol{m} \in \mathbb{R}^p$ and covariance matrix $\Sigma \in S^p_{++}$, where S^p_{++} is the cone of positive definite matrices of $\mathbb{R}^{p\times p}$. For any positive integer $r \leqslant p$, we define the function $\mathcal{M}_{r,\alpha}$ on $\mathbb{R}^p \times S^p_{++}$ by*

$$\mathcal{M}_{r,\alpha} : (\boldsymbol{m}, \Sigma) \mapsto \mathcal{M}_{r,\alpha}(\boldsymbol{m}, \Sigma) = \mathbb{E}_n\left(Z_r^\alpha \ \mathbb{1}_{\boldsymbol{Z}\leqslant\boldsymbol{0}}\right), \tag{4.17}$$

where the inequality $\boldsymbol{Z} \leqslant \boldsymbol{0}$ is to be interpreted component-wise.

Note that the term 'moments of a truncated Gaussian distribution' is technically reserved for $\frac{1}{C_0}\mathcal{M}_{r,\alpha}$, with $C_0 = \mathbb{P}(Z \leqslant \boldsymbol{0})$, but for simplicity we use it here for just $\mathcal{M}_{r,\alpha}$.

If $\boldsymbol{Z}$ is composed of values of a GP at a batch of q locations $\boldsymbol{x}_{n+1:n+q}$, we use the notation $\mathcal{M}_{r,\alpha}(\boldsymbol{Z}_{\boldsymbol{x}_{n+1:n+q}}) \triangleq \mathcal{M}_{r,\alpha}(\boldsymbol{m}(\boldsymbol{x}_{n+1:n+q}), \Sigma(\boldsymbol{x}_{n+1:n+q}))$. We obtain the moments $\mathcal{M}_{r,\alpha}(\boldsymbol{m}, \Sigma)$ of a truncated Gaussian distribution by an extension of Tallis' technique [Tallis, 1961] to any order, presented in the following proposition:

Proposition 10. *The function $\mathcal{G} : \mathbb{R}^p \times \mathbb{R}^p \times S^p_{++} \to \mathbb{R}$ defined by*

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{m}, \Sigma) = e^{\frac{1}{2}\left(\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)^\top \Sigma\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right) - \boldsymbol{m}^\top\Sigma^{-1}\boldsymbol{m}\right)} \Phi_{p,\Sigma}\left(-\boldsymbol{m} - \Sigma\boldsymbol{t}\right), \tag{4.18}$$

where $\Phi_{p,\Sigma}(\cdot)$ is the cumulative distribution function of the centred p-variate normal distribution, is infinitely differentiable, and the moments $\mathcal{M}_{r,\alpha}$ are given by:

$$\mathcal{M}_{r,\alpha}(\boldsymbol{m}, \Sigma) = \left.\frac{\partial^\alpha \mathcal{G}(\cdot, \boldsymbol{m}, \Sigma)}{\partial t_r^\alpha}\right|_{\boldsymbol{t}=\boldsymbol{0}}. \tag{4.19}$$

The proof of this proposition is given in appendix A and relies on calculating the moment generating function $\boldsymbol{t} \to \mathbb{E}\left(\exp\left(\boldsymbol{t}^\top\boldsymbol{Z}\right)\mathbb{1}_{\boldsymbol{Z}\leqslant\boldsymbol{0}}\right)$. Even if an analytical formula can be obtained at any order of differentiation α, the complexity of derivatives in eq. (4.19) increases rapidly. We give below the results for α equals 1 and 2.

Case $\alpha = 1$. Differentiating $\mathcal{G}$ with respect to $\boldsymbol{t}$ yields:

$$\begin{aligned}\frac{\partial\mathcal{G}}{\partial\boldsymbol{t}}(\boldsymbol{t}, \boldsymbol{m}, \Sigma) = &\exp\left(\frac{1}{2}\left(\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)^\top \Sigma\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right) - \boldsymbol{m}^\top\Sigma^{-1}\boldsymbol{m}\right)\right) \times \\ &\left(\Sigma\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)\Phi_{p,\Sigma}\left(-\boldsymbol{m}-\Sigma\boldsymbol{t}\right) - \Sigma\nabla\Phi_{p,\Sigma}\left(-\boldsymbol{m}-\Sigma\boldsymbol{t}\right)\right)\end{aligned}$$

where $\nabla\Phi_{p,\Sigma}$ is the gradient of $\Phi_{p,\Sigma}$ (see appendix A for an analytical derivation). Taking $\boldsymbol{t} = \boldsymbol{0}$ in the previous equation gives

$$\mathcal{M}_{r,1}(\boldsymbol{m}, \Sigma) = m_r\Phi_{p,\Sigma}(-\boldsymbol{m}) - \boldsymbol{\Sigma}_r^\top\nabla\Phi_{p,\Sigma}(-\boldsymbol{m}) \tag{4.20}$$

where $\boldsymbol{\Sigma}_r$ is the r^{th} column of Σ. It is shown in appendix A that computing each of the p components of $\nabla\Phi_{p,\Sigma}$ requires evaluating the CDF of a $p-1$ variate normal distribution. The number of calls to this function for computing the first moment of the truncated Gaussian distribution is thus of $O(p)$.

Case $\alpha = 2$. Similarly, differentiating $\mathcal{G}$ twice with respect to $\boldsymbol{t}$ yields

$$\begin{aligned}\mathcal{M}_{r,2}(\boldsymbol{m}, \Sigma) = (\Sigma_{rr} + m_r^2)\Phi_{p,\Sigma}(-\boldsymbol{m}) + \boldsymbol{\Sigma}_r^\top \ \nabla\nabla^\top\Phi_{p,\Sigma}(-\boldsymbol{m})\boldsymbol{\Sigma}_r \\ + 2m_r\mathcal{M}_{r,1}(\boldsymbol{m}, \Sigma).\end{aligned} \tag{4.21}$$

For readability, the detailed formula of $\nabla\nabla^\top\Phi_{p,\Sigma}$, the Hessian matrix of $\Phi_{p,\Sigma}$, is sent to appendix A. The number of calls to the multivariate normal CDF is of $O(p^2)$.

Calculation of generalised q-EI

The previous results obtained for the moments of the truncated normal distribution turn out to be of interest for computing the generalised q-EI introduced in eq. (4.16), as shown by the following proposition.

Proposition 11. *For $\boldsymbol{x}_{n+1:n+q} \in \mathcal{D}^q$, the criterion* EI_n *defined by eq.* (4.16) *exists for all α and can be written as a sum of moments of truncated normal distributions*

$$\mathrm{EI}_n(\boldsymbol{x}_{n+1:n+q}) = \sum_{\ell=1}^{n}\sum_{r=1}^{q} \mathcal{M}_{n+r-1,\alpha}\left(\boldsymbol{Z}^{(\ell,r)}(\boldsymbol{x}_{n+1:n+q})\right), \tag{4.22}$$

with $\boldsymbol{Z}^{(\ell,r)}(\boldsymbol{x}_{n+1:n+q})$ a vector of size $n+q-1$ defined by

$$Z_i^{(\ell,r)} = \begin{cases} Y_\ell - Y_i & \text{if } 1 \leqslant i \leqslant \ell - 1, \\ Y_\ell - Y_{i+1} & \text{if } \ell \leqslant i \leqslant n-1, \\ Y_r - Y_{i+1} & \text{if } n \leqslant i \leqslant n+q-1 \text{ and } i \neq n+r-1, \\ Y_r - Y_\ell & \text{if } i = n+r-1, \end{cases}$$

noting $Y_i \triangleq Y_{\boldsymbol{x}_i}$.

Moreover, in the noiseless case the random vector $(Y_{\boldsymbol{x}_1}, \dots, Y_{\boldsymbol{x}_n})$ becomes deterministic given $\mathcal{A}_n$. Denoting by ℓ_0 the (smallest) index of the minimal observation, i.e. $Y_{\ell_0} = \min_{\ell=1,\dots,n} Y_\ell$, and writing $\boldsymbol{Z}^{(r)}(\boldsymbol{x}_{n+1:n+q})$ the vector of the q last components of $\boldsymbol{Z}^{(\ell_0,r)}(\boldsymbol{x}_{n+1:n+q})$, Equation eq. (4.22) *is simplified to:*

$$\mathrm{EI}_n(\boldsymbol{x}_{n+1:n+q}) = \sum_{r=1}^{q} \mathcal{M}_{r,\alpha}\left(\boldsymbol{Z}^{(r)}(\boldsymbol{x}_{n+1:n+q})\right). \tag{4.23}$$

***Remark* 2.** *In this thesis we also use the following compact notation for the $(n+q-1)-$dimensional vector $\boldsymbol{Z}^{(\ell,r)}(\boldsymbol{x}_{n+1:n+q})$:*

$$\boldsymbol{Z}^{(\ell,r)}(\boldsymbol{x}_{n+1:n+q}) = A^{(\ell,r)} \left(Y_{\boldsymbol{x}_1}, \ldots, Y_{\boldsymbol{x}_{n+q}}\right)^\top , \tag{4.24}$$

where $A^{(\ell,r)}$ is a matrix implicitly defined by $Z_i^{(\ell,r)}$ of proposition 11.

The proof of proposition 11 is relegated to appendix A for conciseness. Equation (4.22) highlights that the computation of the generalised q-EI in noisy settings is challenging since it involves computing nq different moments, each requiring $(n+q)^\alpha$ calls to the multivariate normal CDF in a dimension close to $n+q$. Even for $\alpha = 1$ and moderate q, linear dependence in the number of observations n makes use of this criterion challenging in application. Regarding the noiseless criterion, the computation of q moments is more affordable, at least for moderate q, but one has to keep in mind that the ultimate goal here is to perform global maximisation of the considered criteria. It is thus important to bring further calculation speed-ups in order to perform this optimisation in a reasonable time compared to the evaluation time of the objective function f, which is assumed to be expensive. The rest of the chapter discusses these matters and proposes faster formulas to compute both q-EI and its gradient.

4.2.2 Speeding up the optimisation of q-EI criterion

Fast numerical estimation of first order moments and their derivatives

Let us now focus on the practical implementation of the closed-form formula eq. (4.22). We take $\alpha = 1$ and note $p = n + q$ in noisy settings and $p = q$ in noiseless settings. As mentioned before, the computation of the noisy or noiseless q-EI (see eqs. (4.22) and (4.23)) requires calls to the CDF of the p-variate and $(p-1)$-variate normal distribution, Φ_p and Φ_{p-1}. These CDFs are here computed using the Fortran algorithms of [Genz, 1992] wrapped in the mnormt R package [Azzalini and Genz, 2014]. A quick look at eqs. (4.20) and (4.22) suggests that the noisy q-EI requires nq evaluations of Φ_p and nq^2 evaluations of Φ_{p-1}. For the noiseless case, the number of calls are divided by n. In both cases, a slight improvement can be obtained by noticing a symmetry which reduces the number of Φ_{p-1} calls from nq^2 (resp. q^2 in the noiseless case) to $nq(q+1)/2$ (resp. $q(q+1)/2$). This symmetry is justified in appendix A.

Despite this improvement, and even in the classical noiseless case, the number of Φ_{p-1} calls is still proportional to q^2. We now give new efficient and trustworthy expansion that enables a fast and reliable approximation of first order moments of truncated Gaussian vectors $\mathcal{M}_{r,1}$ by reducing this number of calls to $O(q)$.

Proposition 12. *Let $\varepsilon > 0$, and let $\boldsymbol{Z}$ be a Gaussian random vector with mean vector and covariance matrix $(\boldsymbol{m}, \Sigma) \in \mathbb{R}^p \times S^p_{++}$. Then we have*

$$\mathcal{M}_{r,1}(\boldsymbol{m}, \Sigma) = \frac{1}{\varepsilon}\left(e^{m_r \varepsilon}\Phi_{p,\Sigma}(-\varepsilon\boldsymbol{\Sigma}_r - \boldsymbol{m}) - \Phi_{p,\Sigma}(-\boldsymbol{m})\right) + O(\varepsilon^2). \tag{4.25}$$

Proof. Let us consider the function $g_r : t \in \mathbb{R} \to e^{m_r t}\Phi_{p,\Sigma}(-\boldsymbol{\Sigma}_r t - \boldsymbol{m})$. This function g_r is tangent at $t = 0$ with the function $t \in \mathbb{R} \to \mathcal{G}(t\boldsymbol{e}_r)$, where the function $\mathcal{G}$ is introduced in proposition 10 and $\boldsymbol{e}_r$ is the r^{th} vector of the canonical basis. It follows from proposition 10 that

$$\mathcal{M}_{r,1}(\boldsymbol{m}, \Sigma) = \left.\frac{\partial \mathcal{G}}{\partial t_r}(\boldsymbol{t}, \boldsymbol{m}, \Sigma)\right|_{t=0},$$

and we obtain the announced result by Taylor expansion of g_r. □

This formula simply uses the approximation of a moment with finite differences of the moment generating function. We showed here that instead of fully computing the moment generating function, we can expand the simpler *tangent* function g_r. For conciseness, we name here the use of this formula as the “tangent moment method”. This formula thus enables approximating the first order moment $\mathcal{M}_{r,1}$ at the cost of only two calls to Φ_p. Hence, from eq. (4.23), computing a noiseless q-EI can be performed at the cost of $2q$ calls to Φ_q. Besides, a similar approach can be applied to approximate the gradient of q-EI through faster computations of $\frac{\partial \mathcal{M}_{r,1}}{\partial \boldsymbol{m}}$ and $\frac{\partial \mathcal{M}_{r,1}}{\partial \Sigma}$, as shown next:

Proposition 13. *The following equations hold:*

$$\frac{\partial \mathcal{M}_{r,1}}{\partial \boldsymbol{m}} = \Phi_{p,\Sigma}(-\boldsymbol{m})\boldsymbol{e}_r - \frac{1}{\varepsilon}\left(e^{m_r\varepsilon}\nabla\Phi_{p,\Sigma}(-\boldsymbol{\Sigma}_r\varepsilon - \boldsymbol{m}) - \nabla\Phi_{p,\Sigma}(-\boldsymbol{m})\right) + O(\varepsilon^2) \tag{4.26}$$

$$\begin{aligned}\frac{\partial \mathcal{M}_{r,1}}{\partial \Sigma} = &-\left(\frac{\partial \Phi_{p,\Sigma}}{\partial x_v}(-\boldsymbol{m})\,\delta_{u,r} + \frac{\partial \Phi_{p,\Sigma}}{\partial x_u}(-\boldsymbol{m})\,\delta_{v,r}\right)_{u,v \leqslant p} \\ &+ \frac{1}{\varepsilon}\left(e^{m_r\varepsilon}\nabla\nabla^\top\Phi_{p,\Sigma}(-\boldsymbol{\Sigma}_r\varepsilon - \boldsymbol{m}) - \nabla\nabla^\top\Phi_{p,\Sigma}(-\boldsymbol{m})\right) + O(\varepsilon^2)\end{aligned} \tag{4.27}$$

where $\nabla\nabla^\top\Phi_{p,\Sigma}$ is the Hessian matrix of $\Phi_{p,\Sigma}$ (see appendix A for details).

As before, these formulas enable reducing the number of calls to the multivariate CDF by an order q. For the computation of q-EI this number goes from $O(q^2)$ to $O(q)$. For computing its dq-dimensional gradient, it goes from $O(q^4)$ to $O(q^3)$. The latter complexity suggests restricting to moderate values of q in applications. In the next section we present results that enable further reducing of the complexity for numerically optimising the q-EI.

Optimising the multipoint Expecting improvement

General observations. Maximizing the EI_n expressions given in eq. (4.22) (noisy settings) or eq. (4.23) (noiseless settings) is difficult. These maximisations are performed with respect to a batch of q points $\boldsymbol{x}_{n+1:n+q} \in (\mathbb{R}^d)^q$, and are thus optimisation problems in dimension dq. In this space, the objective function to be maximised is not convex in general and has the interesting property that the q points in the batch can be permuted without changing the value of EI_n; i.e. $\mathrm{EI}_n((\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q})) = \mathrm{EI}_n((\boldsymbol{x}_{n+\sigma(1)}, \ldots, \boldsymbol{x}_{n+\sigma(q)}))$ for any permutation σ of $\{1, \ldots, q\}$. With this property, one can reduce the measure of the search domain by $q!$, e.g. by imposing that the first coordinate of the q points in the batch are in ascending order. We use here multi-start gradient based local optimisation algorithms acting on the whole input domain $D^q \subset \mathbb{R}^{dq}$, that do not exploit the structure of the problem but do not seem to be affected by this, at least with the chosen settings regarding the starting designs. We propose in section 4.2.2 a faster formula for computing the first moments $\mathcal{M}_{r,1}$ previously presented, as well as their derivatives. This will yield an easier computation of both the generalised EI and its dq-dimensional gradient whose analytical computation is performed in what follows. Besides, a second approximate but faster formula to further reduce the calculation time of the gradient will be introduced in section 4.2.2.

Gradient of the generalised q-EI We provide a calculation of the gradient of q-EI to the case of the generalised noisy and noise-free q-EI. Again, the presented formulas rely on results on moments of truncated Gaussian distributions.

Proposition 14. *Let $\boldsymbol{x}_{n+1:n+q} \in \mathcal{D}^q$ be a batch such that the conditional covariance matrix $\left(\mathrm{cov}\left(Y\left(\boldsymbol{x}_{n+i}\right), Y\left(\boldsymbol{x}_{n+j}\right) \middle| \mathcal{A}_n\right)\right)_{1 \leqslant i,j \leqslant q}$ is positive definite and the functions $\mathbb{E}\left(Y_{\cdot} \middle| \mathcal{A}_n\right)$ and $\left(\mathrm{cov}\left(Y_{\cdot}, Y_{\boldsymbol{x}_{n+j}} \middle| \mathcal{A}_n\right)\right)_{j=1,\ldots,q}$ are differentiable at each point $\boldsymbol{x}_{n+i}$ $(1 \leqslant i \leqslant q)$. These derivatives are written $\boldsymbol{m}'^{(i)} \in \mathbb{R}^d$ and*

$\Sigma'^{(i)} \in \mathbb{R}^{q\times d}$ *respectively. In this setup, the* EI_n *function in eq.* (4.22) *is differentiable and its derivative with respect to the* j^{th} *coordinate of the point* $\boldsymbol{x}_{n+i}$ *is*

$$\frac{\partial \mathrm{EI}}{\partial x_{ij}}(\boldsymbol{x}_{n+1:n+q}) = \sum_{\ell=1}^{n}\sum_{r=1}^{q} m_j'^{(i)} \mathring{A}_i^{(l,r)\top} \frac{\partial \mathcal{M}_{n+r-1,1}}{\partial \boldsymbol{m}}\left(\boldsymbol{Z}^{(\ell,r)}\right) + \tag{4.28}$$

$$\mathrm{tr}\left(A^{(l,r)}\Gamma'^{(i,j)}A^{(l,r)\top}\frac{\partial \mathcal{M}_{n+r-1,1}}{\partial \Sigma}\left(\boldsymbol{Z}^{(\ell,r)}\right)\right),$$

where $\Gamma'^{(i,j)} = \left(\Sigma'^{(i)}_{u,j}\delta_{i,v} + \Sigma'^{(i)}_{v,j}\delta_{i,u}\right)_{u,v} \in \mathbb{R}^{q\times q}$*, and* δ *is the Kronecker symbol. The derivatives* $\frac{\partial \mathcal{M}_{n+r-1,1}}{\partial \boldsymbol{m}}$ *and* $\frac{\partial \mathcal{M}_{n+r-1,1}}{\partial \Sigma}$ *are calculated in appendix A.*

This expansion of the gradient of the generalised EI as a sum of derivatives of first order moments is useful thanks to formulas presented next.

Slightly biased but fast proxy of the gradient

The key idea to obtain further computational savings is summarised in this section. We first strategically decompose the gradient of moments as a sum of two terms.

Proposition 15. *Let us consider a Gaussian multivariate random field* $\boldsymbol{Z} = (Z_i)_{i=1,\dots,p}^\top$ *from* $\mathbb{R}^d$ *to* $\mathbb{R}^p$*. For* $\boldsymbol{x} \in \mathbb{R}^d$*, let us denote by* $\boldsymbol{m}(\boldsymbol{x})$ *and* $\Sigma(\boldsymbol{x})$ *the mean and the covariance matrix of* $\boldsymbol{Z_x}$*. Let* $\boldsymbol{x}_a \in \mathbb{R}^d$ *and assume that* $\Sigma(\boldsymbol{x}_a)$ *is positive definite. Also, assume that the functions* $\boldsymbol{x} \to \boldsymbol{m}(\boldsymbol{x})$*,* $\boldsymbol{x} \to \Sigma(\boldsymbol{x})$ *and* $\boldsymbol{x} \to (\mathrm{cov}(Z_{i,\boldsymbol{x}}, Z_{j,\boldsymbol{x}_a}))_{i,j\leqslant p}$ *are differentiable at* $\boldsymbol{x} = \boldsymbol{x}_a$*. Then the following decomposition holds for* $r = 1,\dots,p$*.*

$$\begin{aligned}&\nabla_{\boldsymbol{x}}\left[\mathcal{M}_{r,\alpha}\left(\boldsymbol{m}(\boldsymbol{x}),\Sigma(\boldsymbol{x})\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a} \triangleq \nabla_{\boldsymbol{x}}\left[\mathbb{E}\left(Z^\alpha_{r,\boldsymbol{x}}\mathbb{1}_{\boldsymbol{Z_x}\leqslant \boldsymbol{0}}\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a}\\ &= \nabla_{\boldsymbol{x}}\left[\mathbb{E}\left(Z^\alpha_{r,\boldsymbol{x}}\mathbb{1}_{\boldsymbol{Z}_{\boldsymbol{x}_a}\leqslant \boldsymbol{0}}\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a} + \nabla_{\boldsymbol{x}}\left[\mathbb{E}\left(Z^\alpha_{r,\boldsymbol{x}_a}\mathbb{1}_{\boldsymbol{Z_x}\leqslant \boldsymbol{0}}\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a}.\end{aligned} \tag{4.29}$$

Proof. $\Sigma(\cdot)$ is continuous at $\boldsymbol{x}_a$, so there exists a neighborhood $V_{\boldsymbol{x}_a}$ of $\boldsymbol{x}_a$ such that for all $\boldsymbol{x} \in V_{\boldsymbol{x}_a}$, $\Sigma(\boldsymbol{x})$ is positive definite. Let us define on $V_{\boldsymbol{x}_a} \times V_{\boldsymbol{x}_a}$:

$$g(\boldsymbol{u},\boldsymbol{v}) = \mathbb{E}\left(Z^\alpha_r(\boldsymbol{u})\mathbb{1}_{\boldsymbol{Z_v}\leqslant \boldsymbol{0}}\right).$$

Applying eq. (A.4) of appendix A, for all $\boldsymbol{u}$ and $\boldsymbol{v}$, $g(\boldsymbol{u},\boldsymbol{v})$ is a moment generated by differentiation of the following function:

$$M_{\boldsymbol{u},\boldsymbol{v}} : t \to e^{\frac{1}{2}\left(\Sigma_{rr}(\boldsymbol{u})t^2 + 2tm_r(\boldsymbol{u})\right)}\Phi_{p,\Sigma(\boldsymbol{v})}\left(-\boldsymbol{m}(\boldsymbol{v}) - t\left(\mathrm{cov}(Z_r(\boldsymbol{u}), Z_j(\boldsymbol{v}))\right)_{j\leqslant p}^\top\right). \tag{4.30}$$

The analytical form of eq. (4.30) and the assumed differentiability at $\boldsymbol{x}_a$ ensure existence of partial derivatives of $g = (\boldsymbol{u}, \boldsymbol{v}) \to \frac{\mathrm{d}^\alpha M_{\boldsymbol{u},\boldsymbol{v}}}{\mathrm{d}t^\alpha}(0)$ at $(\boldsymbol{x}_a, \boldsymbol{x}_a)$. So to conclude,

$$\begin{aligned}\nabla_{\boldsymbol{x}} \left[\mathcal{M}_{r,\alpha}\left(\boldsymbol{m}(\boldsymbol{x}), \Sigma(\boldsymbol{x})\right)\right]\Big|_{\boldsymbol{x}=\boldsymbol{x}_a} &= \nabla_{\boldsymbol{x}} \left[g(\boldsymbol{x}, \boldsymbol{x})\right]\Big|_{\boldsymbol{x}=\boldsymbol{x}_a} \\ &= \frac{\partial}{\partial \boldsymbol{u}} \left[g(\boldsymbol{u}, \boldsymbol{x}_a)\right]\Bigg|_{\boldsymbol{u}=\boldsymbol{x}_a} + \frac{\partial}{\partial \boldsymbol{v}} \left[g(\boldsymbol{x}_a, \boldsymbol{v})\right]\Bigg|_{\boldsymbol{v}=\boldsymbol{x}_a} .\end{aligned}$$

□

The latter decomposition can be interpreted as follows: infinitesimal variations of $(\boldsymbol{m}(\boldsymbol{x}), \Sigma(\boldsymbol{x}))$ around $(\boldsymbol{m}(\boldsymbol{x}_a), \Sigma(\boldsymbol{x}_a))$ modify the moments $\mathcal{M}_{r,\alpha}\left(\boldsymbol{m}(\boldsymbol{x}), \Sigma(\boldsymbol{x})\right)$ in two ways. First, it modifies the distribution of $Z^\alpha_{r,\boldsymbol{x}}$, second it changes the distribution of the truncation $\mathbb{1}_{\boldsymbol{Z_x} \leqslant \mathbf{0}}$. For the particular case of q-EI, we propose to neglect this second variation. Applying this approximation to eq. (4.23) gives for $X_0 \in \mathcal{D}^q$,

$$\begin{aligned}&\nabla_{\boldsymbol{x}_{n+j}} \mathrm{EI}(\boldsymbol{x}_{n+1:n+q})\big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \\ &= \sum_{r=1}^{q} \nabla_{\boldsymbol{x}_{n+j}}\, \mathbb{E}\left(\left(T - Y_{\boldsymbol{x}_{n+r}}\right)^\alpha \mathbb{1}_{A^{(r)} Y_{\boldsymbol{x}_{n+1:n+q}} \leqslant \mathbf{0}}\right)\big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \\ &\approx \sum_{r=1}^{q} \nabla_{\boldsymbol{x}_{n+j}}\, \mathbb{E}\left(\left(T - Y_{\boldsymbol{x}_{n+r}}\right)^\alpha \mathbb{1}_{A^{(r)} Y_{X_0} \leqslant \mathbf{0}}\right)\Big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \\ &= -\nabla_{\boldsymbol{x}_{n+j}}\, \mathbb{E}\left(Y^\alpha_{\boldsymbol{x}_{n+j}} \mathbb{1}_{A^{(j)} Y_{X_0} \leqslant \mathbf{0}}\right)\Big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \\ &= -\mathbb{E}\left(\nabla_{\boldsymbol{x}_{n+j}}\, Y^\alpha_{\boldsymbol{x}_{n+j}}\Big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \mathbb{1}_{A^{(j)} Y_{X_0} \leqslant \mathbf{0}}\right) , \qquad (4.31)\end{aligned}$$

where the last step is obtained by mean square differentiability of the process $\boldsymbol{x} \to {Y_{\boldsymbol{x}}}^\alpha \mathbb{1}_B$, with B an event constant with respect to $\boldsymbol{x}$, see appendix A. We can observe that this approximation makes a summation term disappear. The computation of this formula requires $(d+1)$ evaluations of q-variate Gaussian CDF. Indeed, eq. (4.31) indicates that each component of the gradient vector can be considered as a moment of a truncated Gaussian vector, so we can apply the results of section 4.2.1. In particular, when $\alpha = 1$, applying proposition 12, two Gaussian CDF calls are needed for each of the d components, leading to $2d$ evaluations. Besides, from eq. (4.25), the second CDF call does not depend on r, which implies that this term is common for every dimension. Thus the gradient of eq. (4.31) finally comes with $d+1$ CDF evaluations instead of $2d$. For a full gradient with respect to all q points of the batch, we then need $q(d+1)$ CDF evaluations – a substantial improvement compared to the $O(q^4)$

obtained in [Marmin et al., 2015] and the $O(q^3)$ obtained in the previous section. The complexities for computing moments, q-EI and its gradients, expressed in terms of number of calls to the Φ function, are summarised in table 4.1. These new computational savings come at the price of a non-exact gradient calculation. A first numerical validation is represented in fig. 4.4. On this example, we observe small (1×10^{-2}) relative errors between the exact and approximate gradient of dimension $q \times d = 4$ (the biggest difference vector has a norm of 0.13, compared to an exact gradient norm of 13.1). We also observe that the relative error appears to be typically smaller with higher q-EI, which is promising for q-EI maximisations. However, this apparently trustworthy but non-exact calculation naturally raises the question of the impact of such an approximation on the performance of gradient-based q-EI maximisation algorithms. As we will see in the next section, this proxy gradient turned out to enable quite competitive q-EI maximisation performances based on numerical experiments.

Table 4.1: In noiseless settings, total number of calls to the CDF of the multivariate Gaussian distribution for computing $\mathcal{M}_{r,1}$, q-EI, their gradients and their approximations, depending on q and d. For q-EI in noisy setting, replace q by $p = n + q$ and multiply each number of calls by n.

		Number of CDF evaluations				
		Φ_{q-3}	Φ_{q-2}	Φ_{q-1}	Φ_q	Total
$\mathcal{M}_{r,1}$	analytic			q	1	$O(q)$
	tangent moment				2	2
EI	analytic			$\binom{q+1}{2}$	q	$O(q^2)$
	tangent moment				$2q$	$O(q)$
$\nabla\mathcal{M}_{r,1}$	analytic	$3\binom{q}{3}$	$3\binom{q}{2}$	$2q$	1	$O(q^3)$
	tangent moment		$2\binom{q}{2}$	$2q$	2	$O(q^2)$
	proxy				$d+1$	$O(d)$
∇EI	analytic	$6\binom{q+1}{4}$	$3\binom{q+1}{3}$	$(3q^2+q)/2$	q	$O(q^4)$
	tangent moment		$q^2(q-1)$	$2q^2$	$2q$	$O(q^3)$
	proxy				$q(d+1)$	$O(qd)$

Calculation speed

Here we illustrate the usability of the proposed gradient-based q-EI maximisation schemes and in particular the improvements brought by the fast formulas detailed in the previous sections. The relevance of using sequential sampling

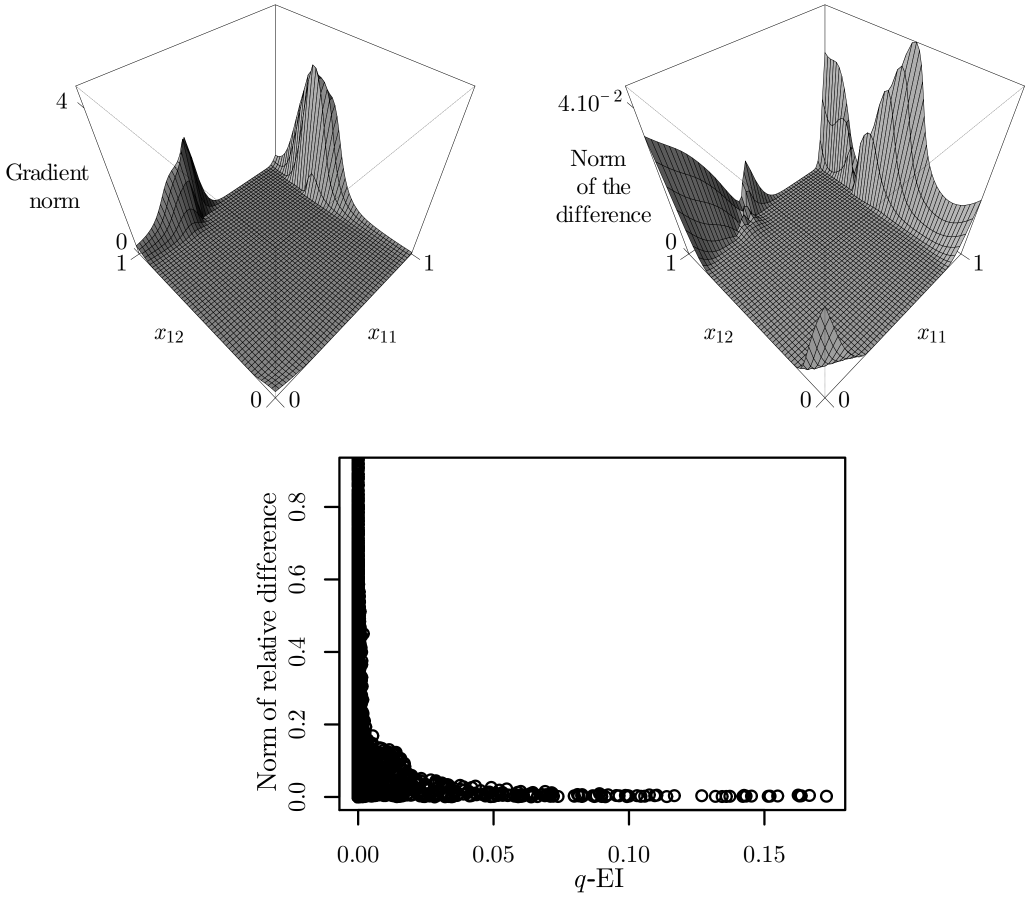


Figure 4.4: Numerical validation of the approximation from eq. (4.31), with $\alpha = 1$, $q = 2$, $d = 2$. From left to right: 1) Norm of the q-EI gradient, with respect to the first batch point (the other point is fixed in the center of $[0,1]^d$) ; 2) Norm of the difference vector between the analytical gradient and its approximation ; 3) Relative error (norm of the difference divided by the real norm) computed on 3000 random batches sampled uniformly in $[0,1]^{d\times q}$, with respect to their q-EI.

strategies based on the q-EI maximisation has already been investigated (see, [Chevalier and Ginsbourger, 2014., Wang et al., 2015, Marmin et al., 2015]) and all these articles pointed out the importance of calculation speed which often limits the use of q-EI based strategies to moderate q. We do not aim again at proving the performance of q-EI based sequential strategies. Instead we aim at illustrating the gain, in computation time, brought by the fast formulas and show that using the approximate gradient obtained in eq. (4.31) does not impair the ability to find batches with (close to) maximal q-EI.

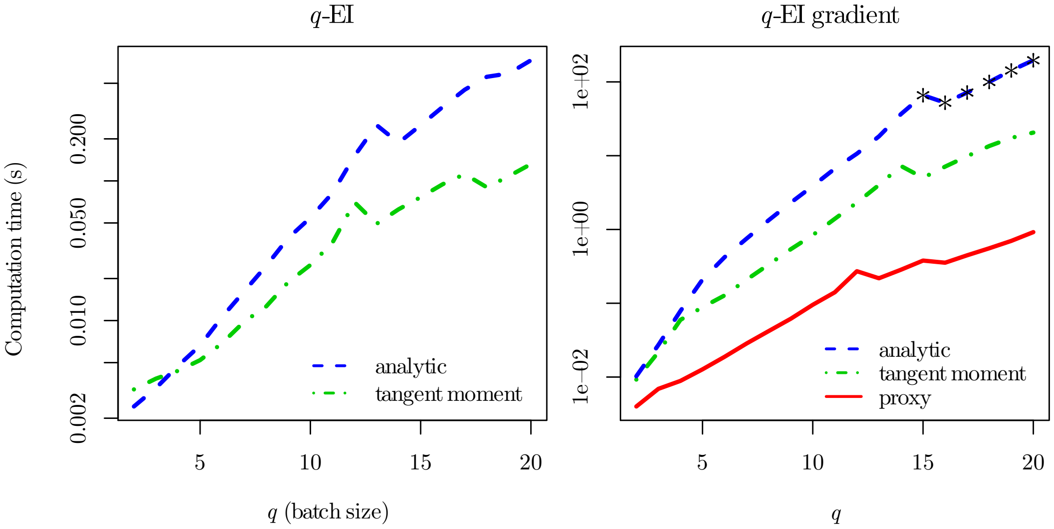


Figure 4.5: Computation times for q-EI or its gradient as a function of the batch size q (logarithmic scale). We take an averaged computation time over 1000 batches (except for points marked with a $*$, averaged over 150 batches).

In chapter 5, we apply sequential strategies to minimise f. Here we look at empirical computation times for evaluating q-EI and its gradient as a function of the batch size q. For the computations, the so-called "analytic" method relies on the state of the art formulas of [Chevalier and Ginsbourger, 2014., Marmin et al., 2015] with a number of calls to the multivariate normal CDF of respectively $O(q^2)$ and $O(q^4)$. The "tangent moment" method uses our formula for moment calculation to yield q-EI and its gradient (see eqs. (4.25) to (4.27)). Finally, for computing the gradient only, the "proxy" method relies on eq. (4.31).

Figure 4.5 exhibits computation times averaged over 1000 batches drawn uniformly. The Gaussian process model is based on an initial design of $n_0 = 10d = 80$ points drawn from an optimum-LHS procedure [Kenny et al., 2000]. We use the Matérn ($\nu = 3/2$) tensor product covariance function and estimate the hyperparameters by maximum likelihood using the DiceKriging R package [D. Ginsbourger and V. Picheny and O. Roustant and with contributions by C. Chevalier and S. Marmin and T. Wagner, 2015]. Figure 4.5 shows significant computational savings. For instance with $q = 8$, one gradient computation takes respectively $0.04s$, $0.33s$ and $1.33s$ using respectively the proxy, the tangent moment and analytic methods. Since the complexity for computing a gradient with the proxy is of $O(qd)$ against $O(q^3)$ and $O(q^4)$ for the two other methods, the computational savings of the proxy tends to increase with q. It should also be noted that these savings will be larger with decreasing domain dimension d. If we look at q-EI computations, the tangent moment method is 3.3 times faster than the analytic one when $q = 8$ and 6.5 times faster when $q = 20$; thanks to an $O(q)$ complexity against $O(q^2)$.

In this second part of the chapter, we have provided a closed-form expression of generalised q-points Expected Improvement criterion for batch-sequential Bayesian global optimisation. An interpretation based on moments of truncated Gaussian vectors yields fast q-EI formulas with arbitrary precision. Furthermore a new approximation for the gradient is shown to be even faster while preserving ability to find batches close to maximal q-EI. As the use of these strategies was previously considered cumbersome from a dozen of batch points, these formulas happen to be of particular interest to run q-EI based batch-sequential strategies for larger batch sizes. Additionally, some of the intermediate results established here might be of interest for other research questions involving moments of truncated Gaussian vectors and their gradients. In section 5.4, we apply these methods on a classic 8-dimensional test case. In particular, a multistart derivative-based multipoint EI maximisation algorithm highlighting the benefits of the considered methodological principles and the proposed fast approximations is tested and compared to baseline strategies.

Chapter 5

Numerical experiments on engineering applications

This chapter deals with a series of applications in engineering where special attention is devoted to the assessment of the capabilities of the methodological contributions developed in this thesis for the approximation of functions with heterogeneous variations. Also, we present numerical experiments pertaining to speed-ups in the computation of the batch-sequential generalised expected improvement criteria and its gradient, for parallel global optimisation.

Our numerical experiments include several performance benchmarks, be it in terms of sampling criteria (ranging from MSE, IMSE to GNV, IGNV and also EI) or surrogate models (stationary anisotropic, TGP [Gramacy, 2007, Gramacy and Taddy, 2010], CGP [Ba and Joseph, 2014], WaMI-GP). In this chapter, experiments are in R [R Core Team, 2015]. In all considered experiments, initial experimental designs are based on Latin Hypercube Sampling optimised with maximin distance (see section 2.1.2) from the package DiceDesign [Dupuy et al., 2015].

The baseline model throughout the chapter is ordinary kriging in empirical Bayes settings. i.e. the constant mean and the covariance function of the underlying GP are supposed to be known, and the covariance parameters are estimated by maximum likelihood using the ‘mle’ function of the kergp R package [Deville et al., 2015]. More generally the topic of trends is not addressed further in the considered approaches, for simplicity. For the gradient-based criteria, the code computing the conditioned distribution of the GP gradient is programmed in C++, embedded in R using RcppArmadillo, an algebra package [Eddelbuettel and Sanderson, 2014].

Except for optimisation, the capability assessment is achieved by focusing on estimates of the L^2 prediction error:

$$\Delta_N = \left(\int_{\mathcal{D}} (m_N(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \right)^{1/2} , \tag{5.1}$$

with m_N the predictor obtained by the tested method after a total budget of N evaluations has been spent. Experimental design strategies are replicated by starting from different initial designs in order to account for stochastic effects.

The next section is a brief description of the context of the applications. Then, we provide several numerical benchmarks devoted to the approximation of functions with heterogeneous variations (Sections 5.2 and 5.3) and to global optimisation (Section 5.4). In all the tests, the input variables are rescaled between 0 and 1 and are denoted by $x_1, \ldots, x_d$ for the sake of clarity.

5.1 Presentation of case studies

From an applied perspective, the thesis is motivated by engineering problems, and in particular by IRSN test cases in mechanical engineering for civil nuclear safety. The overall goal is to enable an enhanced analysis of systems with heterogeneous variations in contexts, such as those of expensive numerical simulations, where the number of evaluations is limited. The main test cases are taken from a mechanical simulator of cracking propagation in a material. Further numerical comparisons are performed on a fluid dynamics test case from NASA. For our contributions in optimisation, we exploit a well known deterministic function that models waterflow through a borehole. In the following sections we detail the context of each of these applications.

5.1.1 Cracking simulation of composite materials

The context is here the mechanical study of nuclear installations. Cracking propagation in components due to ageing of nuclear plants is a safety issue studied at IRSN and in particular in the Micromechanics and Structural Integrity Laboratory, a joined laboratory of IRSN and the National Center for Scientific Research (CNRS). One important goal is to understand the links between the characteristics of the materials constituting the components and accidental radioactive leaks. The cracking energy of a component, the smallest energy required to break the material apart, is a key value for risk control. It is the output of interest in the following two applications. The cracking energy with respect to some mechanical parameters of the component (detailed in what follows), as well as a images of different cracking propagations are displayed in figs. 5.1 and 5.3.

The cracking is simulated using a computer program *Xper* [Perales et al., 2010] (using the solving method 'NonSmooth Contact Dynamics' [Perales et al., 2008]). In the simulations, components in undamaged state are submitted to a specified force until they are entirely torn apart.

The numerical techniques for the simulation are precise but costly: the average computation time for one evaluation is 2 to 3 days[1]. Depending on the inputs, the duration can even reach one week. The goal of these applications is to obtain a description as precise as possible of the relation between some input parameters of the Xper code and the cracking energy in output. Because of the limited computational budget, there is a high interest in choosing carefully the evaluations with model-based design of experiments in order to capture the physical behaviour.

We briefly describe now the mechanical properties of the considered material. For a detailed version, Perales et al. [2008] describes this structure named 'MP-CZM'. The structure consists of squares with side length 2.8×10^{-2} m periodically repeated over space in two orthogonal directions. The traction force applied to the structure is uniform, axial and of intensity 10^2 s^{-1}. A cracking seed (determining where the propagation starts) is positioned in the middle of the left side. The composite material has two phases: the matrix, and inside of it, the inclusions[2]. Figure 5.2 displays the matrix/inclusion layout identical in all squares.

The matrix and the inclusion are supposed to be elastic. The cohesion at a given interface between the matrix and the inclusion is described by a ratio $w^{mat-incl}/w^{incl-incl}$ (varying here between 1.3×10^{-3} and 7.5×10^{2}). Referring to [Perales et al., 2008], we mention that $w^{mat-incl}$ is named 'surface energy of the interface matrix/inclusion' (also interface energy) and $w^{incl-incl}$ is the surface energy of the interface inclusion/inclusion (also inclusion energy). The lengths of the inclusions can have also some impact on the behaviour of the output. They vary here between 3×10^{-3} and 1×10^{-2} m.

[1] Time experienced on a processor Xeon Intel with frequency 3.20 GHz and 8 GB of RAM).

[2] A matrix of a composite material is a binding agent (of ceramic, plastic, metal, etc) that holds other materials together (here named 'inclusions'), forming a more complex solid material overall.

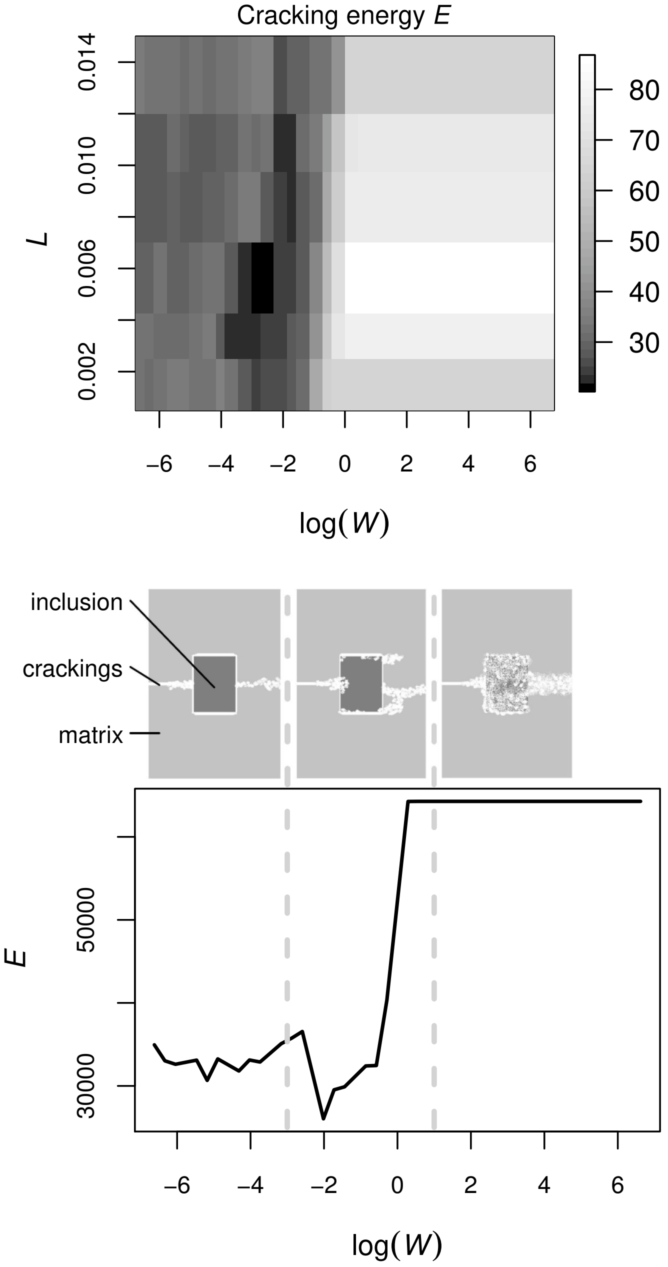


Figure 5.1: Cracking energy of an heterogeneous material depending on two uncertain input parameters (ln(W) and L). The image on the top represents the available dataset while the image on the bottom is a 1D-cut along the ln(W)-axis with sketches of the cracking phenomena. According to the inputs, the cracking propagates around or through the inclusion. These two modes correspond respectively to high or low cracking energies. A transition zone appears in between with a steep slope.

Two test cases are studied depending on the number of inclusions in the matrix. They are referred as ‘test case 1’ and ‘test case 2’.

Test case 1

In this test, the matrix contains one inclusion (at the center of the square) (fig. 5.1) and we have two uncertain input parameters:

- the ratio $W = w^{mat-incl}/w^{incl-incl}$,
- the length L (along the y axis).

The available dataset to evaluate the capability of each method includes 216 points corresponding to the simulation of the cracking energy on a 36×6 grid, see fig. 5.1. One can see that a high variation zone located along a straight line, slightly non-aligned with the canonical axes. We also show on this figure a unidimensional cut of the cracking energy with respect to $\ln(W)$. We observe a region where a small variation of the inputs impacts drastically the output. This can be is attributed to the competition between different physical phenomena that control the trajectory of the cracking (it can go through the inclusion or around it).

Test case 2

In this test, the matrix has two inclusions (fig. 5.2) and we have four uncertain input parameters :

- for the first inclusion, the ratio $W_1 = w^{mat-incl}/w^{incl-incl}$,
- this same ratio W_2 for the second interface,
- the length L_1 (in the y axis, see fig. 5.2),
- this same length for the second inclusion L_2.

Figure 5.3 shows a bidimensional slice of the cracking energy with respect to the interface ratios W_1 and W_2, with fixed inclusion lengths $L_1 = 1 \times 10^{-2}$ m and $L_2 = 6 \times 10^{-3}$ m. Similar to the first test case, the plot exhibits several zones where the output has high variations. However, these transition zones are more complex and therefore more difficult to capture by the tested methods.

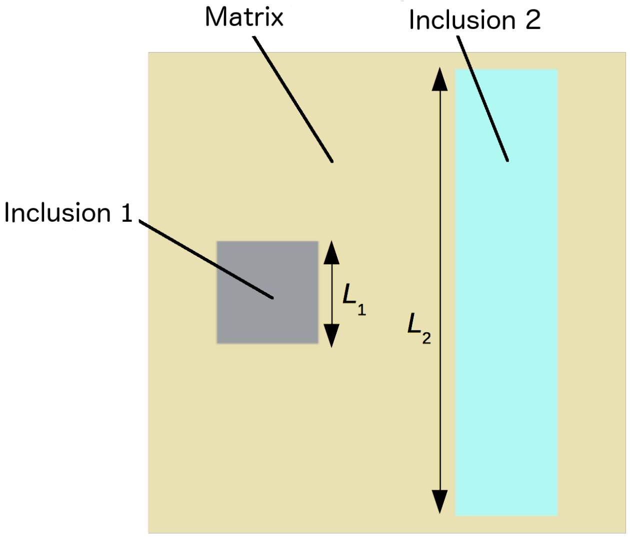


Figure 5.2: Scheme of the structure of the composite material for the IRSN test case 2.

The numerical comparisons are performed on the bivariate case of fig. 5.3. The data set includes 289 evaluations, enabling a computation of the prediction error Δ_N. The evaluations were designed with Latin hypercube sampling (LHS) optimised with the maximin criterion (see section 2.1.2).

5.1.2 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 Rogers et al. [2003]. 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 set is displayed in fig. 5.4. We see that variations are this time mainly directed by canonical axes. 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 $x_1 \approx 0.5, x_2 \geqslant 0.5, x_3 \leqslant 0.5$). These are due to the complexity of the simulator whose convergence depends on a solver which sometimes returns inaccurate values despite automatic checks [Gramacy and

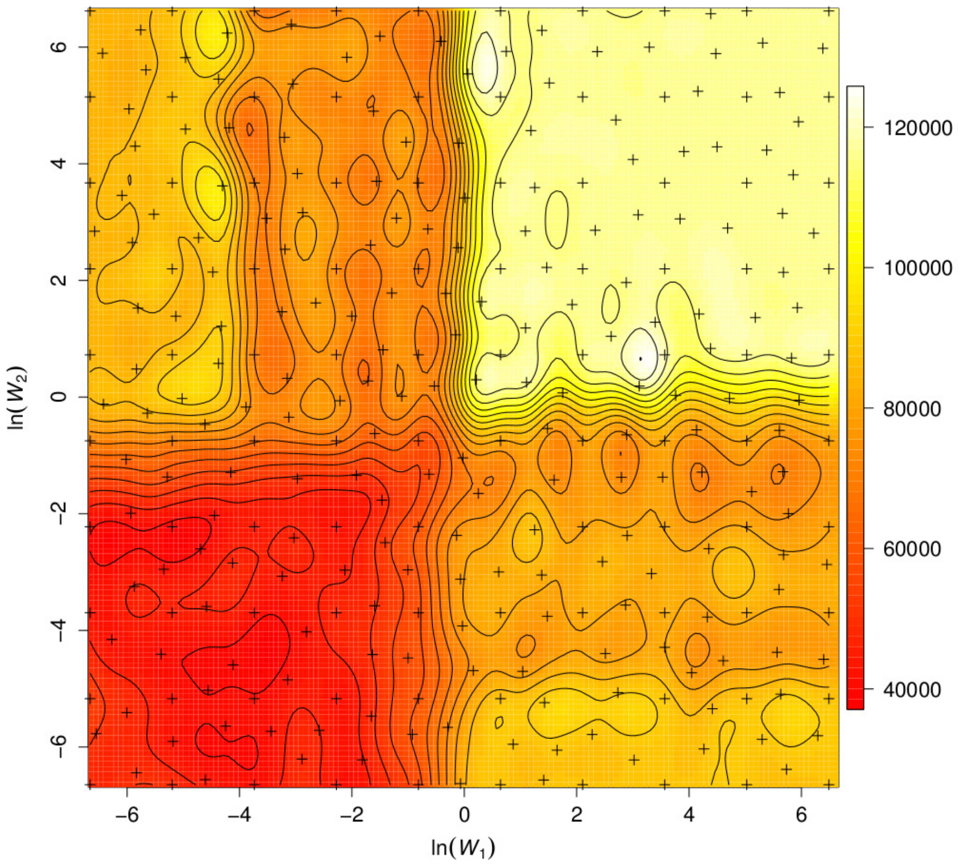


Figure 5.3: Cracking energy on a bidimensional slice of the input space (length parameters fixed to $L_1 = 1 \times 10^{-2}$ m and $L_2 = 6 \times 10^{-3}$ m.), with respect to the logarithms of the energy ratios.

Lee, 2008]. Thus the models will be considered in noisy setting in order to smooth out the convergence errors.

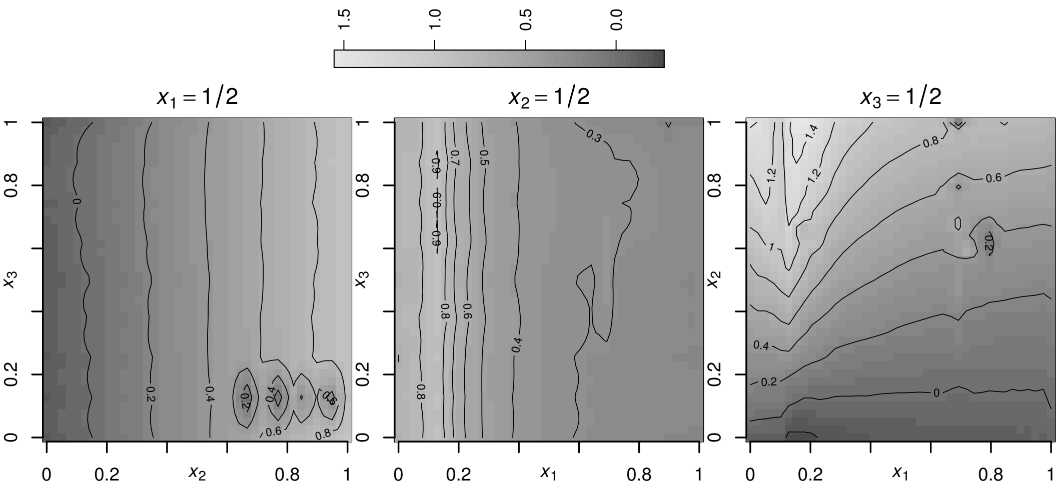


Figure 5.4: Simple tridimensional GP interpolation of the available data on the Langley Glide-Back Booster simulation test case (only 3 slices of the input cube are displayed). The interpolation is made with the R package DiceKriging [Roustant et al., 2012] with default parameters (ordinary kriging, MLE, with tensor product Matérn kernel, smoothness $\nu = 5/2$).

5.1.3 Borehole function

The Borehole function [Harper and Gupta, 1983] has been previously used for testing methods using a surrogate model [Worley, 1987, Gramacy and Lian, 2012a]. The function computes a rate of water flow, ϕ, through a borehole. The problem is described by $d = 8$ input variables, $r_w \in [0.05, 0.15]$, $r \in [100, 50000]$, $T_u \in [63070, 115600]$, $H_u \in [990, 1110]$, $T_l \in [63.1, 116]$, $H_l \in [700, 820]$, $L \in [1120, 1680]$, $K_w \in [1500, 15000]$ and is given below

$$\phi = \frac{2\pi T_u (H_u - H_l)}{\ln\left(\frac{r}{r_w}\right)\left(1 + \frac{2LT_u}{\left(\ln\left(\frac{r}{r_w}\right) r_w^2 K_w\right)} + \frac{T_u}{T_l}\right)}. \tag{5.2}$$

Here, the objective function f is obtained by rescaling ϕ on the input domain $\mathcal{D} = [0, 1]^8$. An analytical study of variations shows that there is a unique global minimum at $\boldsymbol{x}^* = (0, 1, 0, 0, 0, 1, 1, 0)^\top$, with $f(\boldsymbol{x}^*) \approx 1.1918$.

5.2 Comparisons of models and sampling criteria

5.2.1 Comparisons of predictions from fixed space-filling designs

We first compare the predictive performances of stationary GP, WaMI-GP and TGP methods.

• IRSN test case 1

We built 5000 space-filling designs of size 20 (a set of LHS designs optimised with a maximin criterion see e.g. [Dupuy et al., 2015]). 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 γ_1 and γ_2 the standard parametrisation with cumulative distribution functions of beta distributions described in section 3.1. The results displayed on fig. 5.5 indicate that our approach outperforms the two other ones in terms of prediction errors.

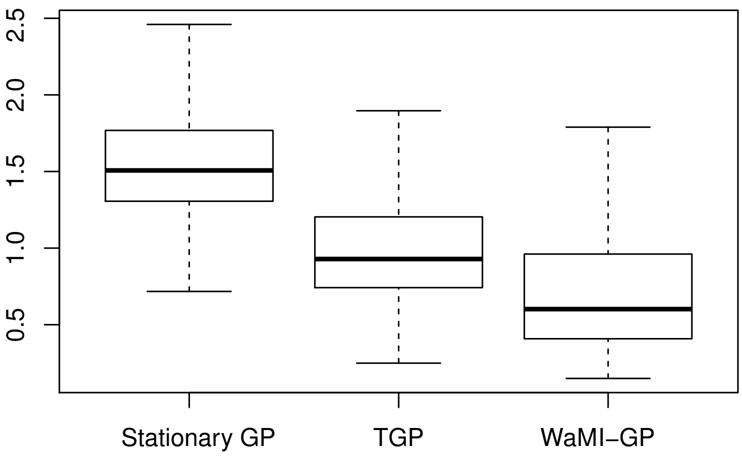


Figure 5.5: Comparison of L^2 prediction errors on the IRSN test case between the three candidate models: stationary anisotropic GP, TGP and WaMI-GP. The boxplots are obtained from repetitions with 5000 different initial designs.

It is also informative to analyse the estimated (overall) warpings, as illustrated in fig. 5.6 (we take the warping from the design giving a median prediction error). 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 WaMI-GP model allows linear transformation of the input space, these lines do not have to be aligned with canonical axes, adapting with more freedom their directions to the shape of the actual high variation region.

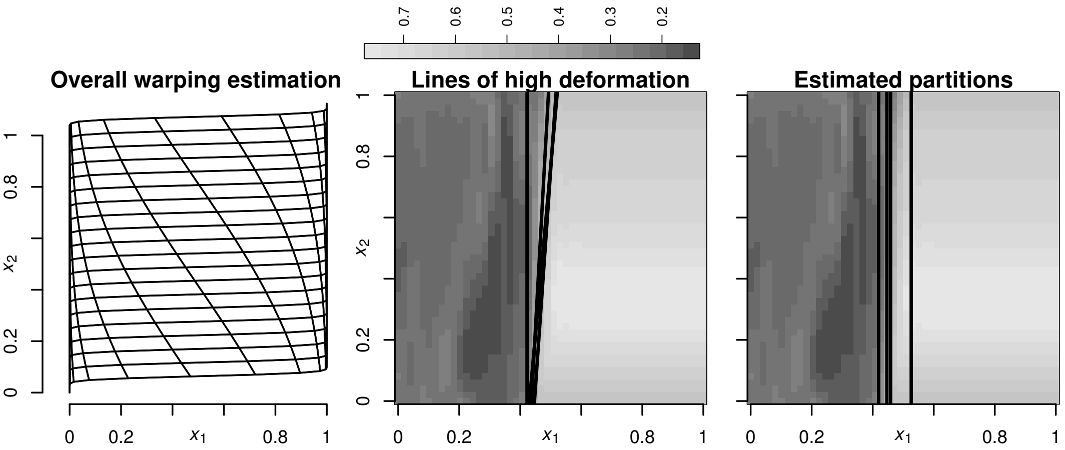


Figure 5.6: Some features of models with median prediction errors. Left: estimated warping of the WaMI-GP model with median predictivity; middle: lines of maximal distortion for 5 (most) median models; right: lines of partitioning for 5 median TGP models.

- Langley Glide-Back Booster (LGBB)

In order to study the influence of the initial number of evaluations, several tests are performed for designs including from 50 to 700 points (maximin-optimised LHS). All experiments are repeated 50 times with a different initial designs[3]. We then focus on the median and 95% quantile of the errors. Table 5.1 provides the prediction error associated with our WaMI-GP approach versus TGP.

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) according to the data while in its present form WaMI-GP has a fixed structure prescribed

[3] The differences correspond to different values of the seeds of the algorithm generating the LHS, see [Dupuy et al., 2015].

Table 5.1: Prediction errors of the models on the LGBB data.

	$N_0 = 50$		$N_0 = 100$		$N_0 = 150$		$N_0 = 300$		$N_0 = 700$	
	$q_{50\%}$	$q_{95\%}$	$q_{50\%}$	$q_{95\%}$	$q_{50\%}$	$q_{95\%}$	$q_{50\%}$	$q_{95\%}$	$q_{50\%}$	$q_{95\%}$
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

by the user. However in the following section, we will see that WaMI-GP may outperform TGP when adding experiments based on the MSE criterion.

5.2.2 Comparisons of model-based sequential designs

We now investigate the capability of the proposed criteria in *sequential* design settings. For different criteria (among MSE/IMSE and $\mathrm{GNV}_{\eta=1,2}/\mathrm{IGNV}_{\eta=1,2}$), we repeat several steps of the sequential design: point selections coupled with re-estimation of the model parameters. A simplified version of $\mathrm{IGNV}_{\eta=1,2}$ is also considered by plug-in of the mean value in the integrand. More precisely,

$$J_n^{\overset{\text{IGNV}}{\text{plug-in}},\eta}(\boldsymbol{x}) = \int_{\boldsymbol{u}\in D} \mathrm{var}\left(||\nabla Y_{\boldsymbol{u}}||^{\eta}|\,\mathcal{A}_n, Y_{\boldsymbol{x}} = m_n(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{u}. \tag{5.3}$$

- IRSN test case 1

With the very high computation cost of simulation runs in mind, ten new evaluations are added starting from a space-filling design of $n = 20$ points. The tested GP models are stationary isotropic and WaMI. The whole workflow is replicated 100 times and the results are displayed in fig. 5.7 and table 5.2.

Let us first notice that the WaMI-GP model generally leads to the smallest prediction errors. This can reasonably be attributed to an adaptation of the model to the function f exhibiting a steep transition region.

When the model is stationary, the MSE and IMSE criteria do not focus on adding points in the steep transition region (one can say these methods explore D in a space-filling way). On the contrary, the $\mathrm{IGNV}_{\eta=1}$ criterion detects regions where the gradient's norm is high, leading to a better model training and to a 50% reduction 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

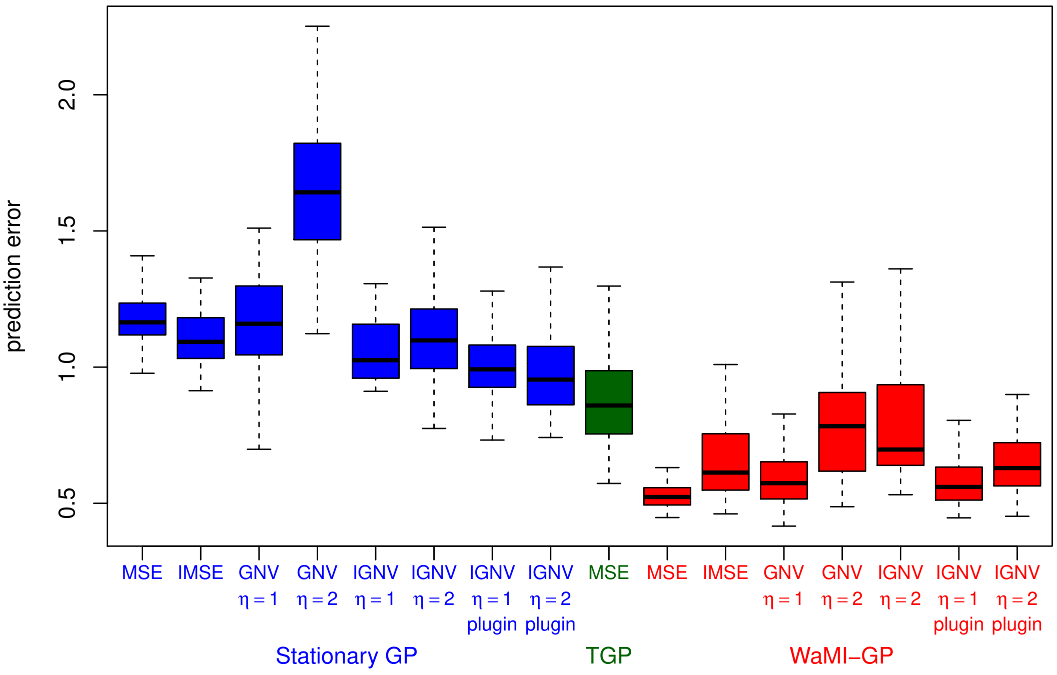


Figure 5.7: Distribution of the prediction error for the 'IRSN test case 1' after different sequential design of experiments. Sampling criteria are compared in both stationary and WaMI-GP models.

Table 5.2: For IRSN test case 1, required number of steps for achieving 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

30% of the number of simulations compared to the stationary framework. Finally, coupling WaMI-GP modelling and gradient-based criteria leads to rather poor results on fig. 5.7 since, by construction, both aspects contribute to more exploitation in targeted regions and their effects add up, with detrimental consequences on the exploration side.

- Langley Glide-Back Booster

From initial designs of size 50 (still using LHS with optimised maximin distance), we perform 20 new evaluations chosen by MSE maximisation. Results obtained in prediction with TGP and the WaMI-GP model are presented in fig. 5.8. We see that the prediction errors are reduced faster using the WaMi-

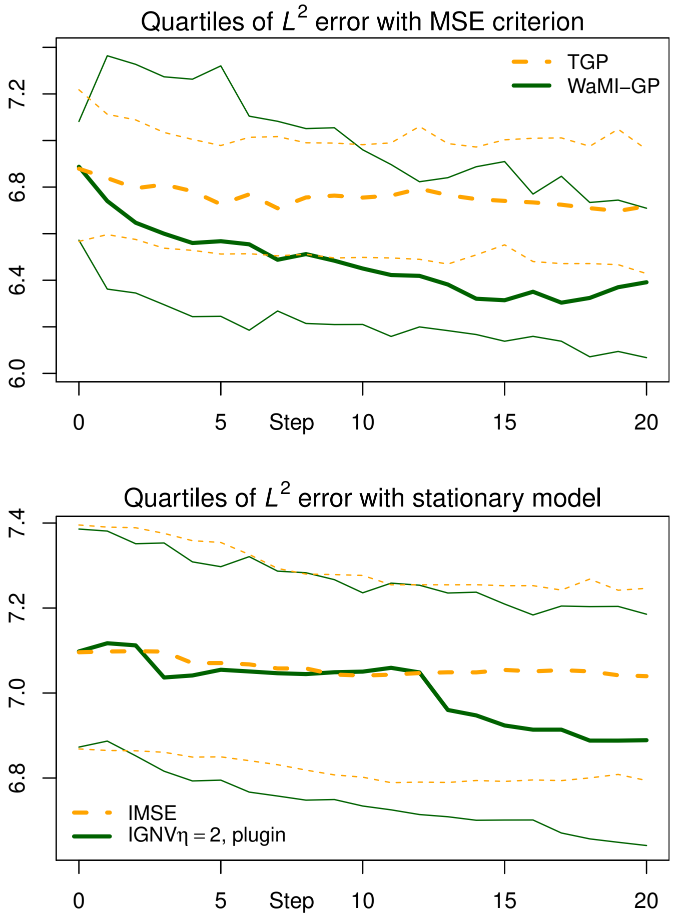


Figure 5.8: Medians (plain) and quartiles (dotted) of prediction errors during sequential designs of experiments. Top: comparison of models TGP and WaMI-GP with a common criterion MSE. Bottom: comparison of criterion IMSE and IGNV, $\eta = 2$, with a standard stationary GP model.

GP model. Indeed, the estimated warping allows to dilate the input space in the region of high variations (around Mach 1). It results in an increased model variance in this area and thus a more dense exploration of it via MSE maximisation. 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. [Gramacy and Lee, 2009], where similar variance-based criteria, Active Learning-MacKay and Active Learning-Cohn, are used for asynchronous batch sequential applications). We also compare again a gradient-based criterion, IGNV, $\eta = 2$ with a classical criterion IMSE relying on a stationary anisotropic GP model (fig. 5.8).

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.

5.3 Further experiments on the approximation of functions with heterogeneous variations

We consider in this section some extensions of designs of experiments based on input space warping, and evaluate their capability on the previous test cases. The first extension integrates a WaMI-based multipoint strategy to add a batch of evaluations in order to launch several simulations on a computer cluster. The second one is the Wav-GP approach proposed in section 3.4 that allows a non-parametric estimation of the warping of a non-stationary GP model.

5.3.1 WaMi-GP and multipoint sampling

WaMI-GP is combined with a multipoint version of the MSE criterion defined in terms of the determinant of the posterior covariance matrix (see section 2.1.2). All the comparisons are performed on the second IRSN test case.

Gain of using a sequential design

Since this test case has not been studied in the previous section, we first consider the sequential approach with batchsize $q = 1$ and compare with fix space-filling designs with sizes from 10 to 40.

The sequential design starts with $n_0 = 10$ initial points. Here as in the rest of the section, the experiments are repeated 50 times with initial and non-adaptive (fixed) designs drawn randomly (optimised LHS).

We display on fig. 5.9 the median prediction error with respect to the number of evaluations or to the evaluation time. These experiments are made twice: once with the stationary anisotropic model and once with WaMI-GP model.

With a same non-stationary model, the sequential design leads to a more accurate prediction than the non-adaptive LHS design. Indeed fig. 5.10 shows that the warping has dilated the areas of higher variation, i.e. the zones around the two lines of equations $x_1 = 0$ and $x_2 = 0$, where the function has high variation. Thanks to the detection of zones with high variations by the model, the sequential method evaluates more intensely in these zones, and improves the predictions. This effect is not experienced with a stationary model, where the sequential method does not perform better than a (simpler) model-independent method.

Stationary and non-stationary models can be also compared in term of number of evaluations to reach a given error. It is displayed by fig. 5.11 and shows that using the non-stationary model saves up to 55% of evaluations.

Numerical gain of multipoint sequential design

Here we use the multipoint version of the MSE described in section 2.1.2 to sample new evaluations.

Figure 5.9 also gives the prediction error with respect to the computation time when increasing batchsize $q = 1$, 5 and 10. Obviously the computation time for outperforming a given accuracy is much shorter when $q > 1$. A better indicator of the efficiency of parallelisation is the *speedup*, defined as t_q/t_1, with t_i the computation time of the case $i = q$ (i.e. how many times the computation is faster than the case $q = 1$). Figure 5.12 displays this value for different prediction error levels and batchsizes.

We observe that the speedup is higher for lower error levels. For example, in the stationary case, and for an error of 1.95, increasing q from 1 to 10 speeds

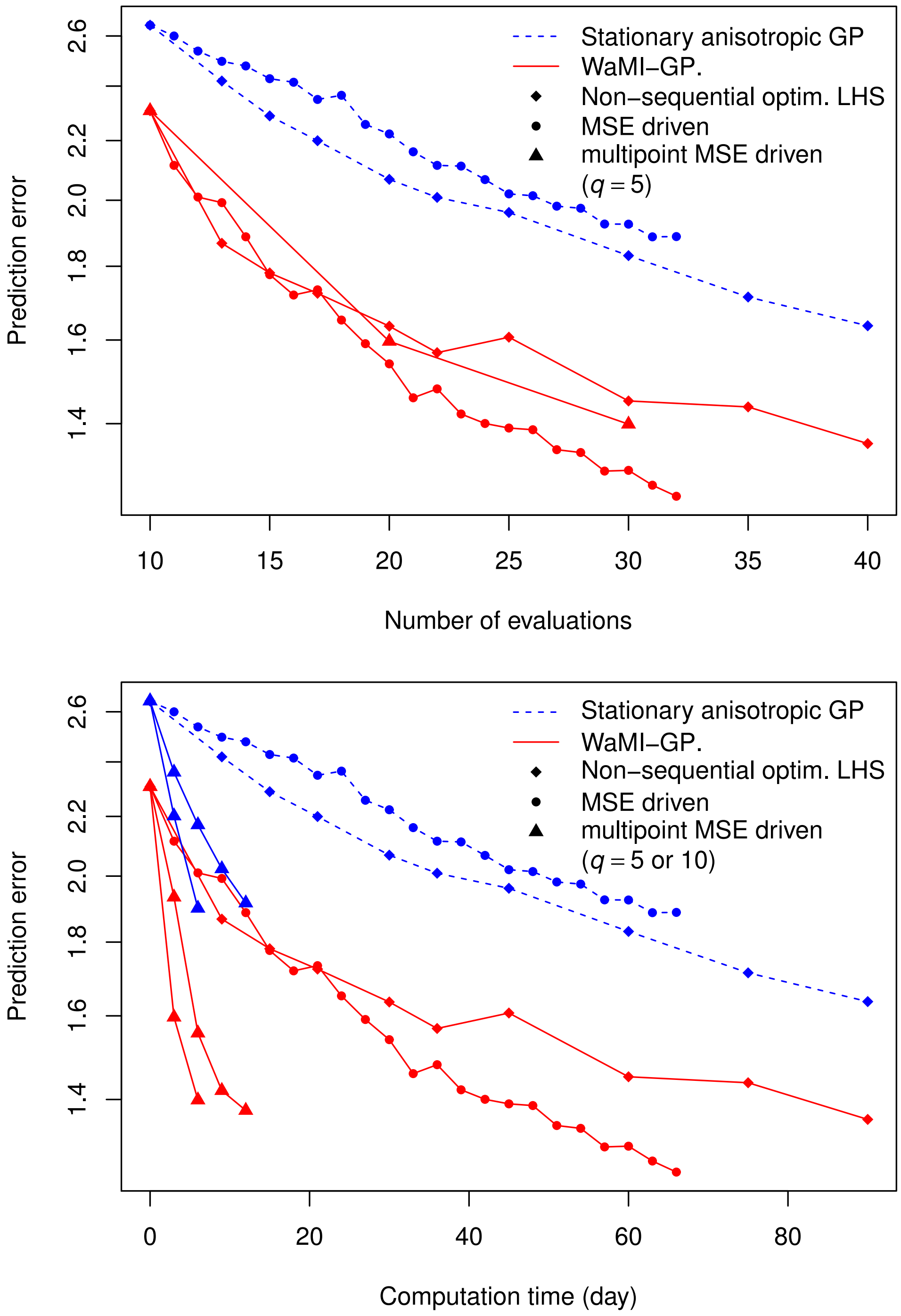


Figure 5.9: Prediction error with respect to the number of evaluations or the computation time for different methods of design of experiments.

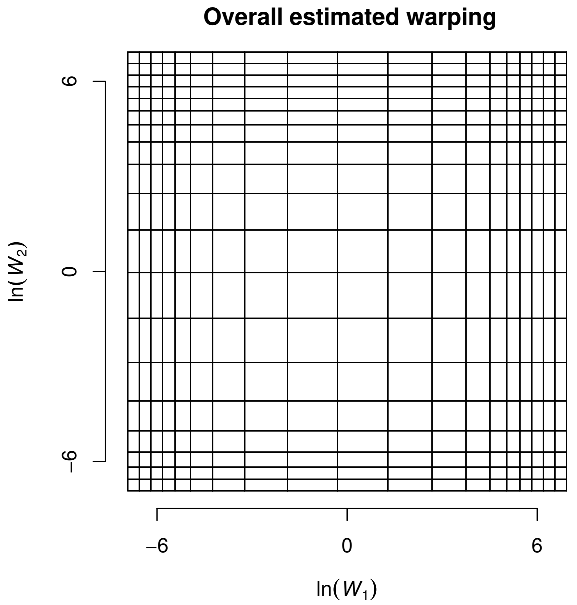
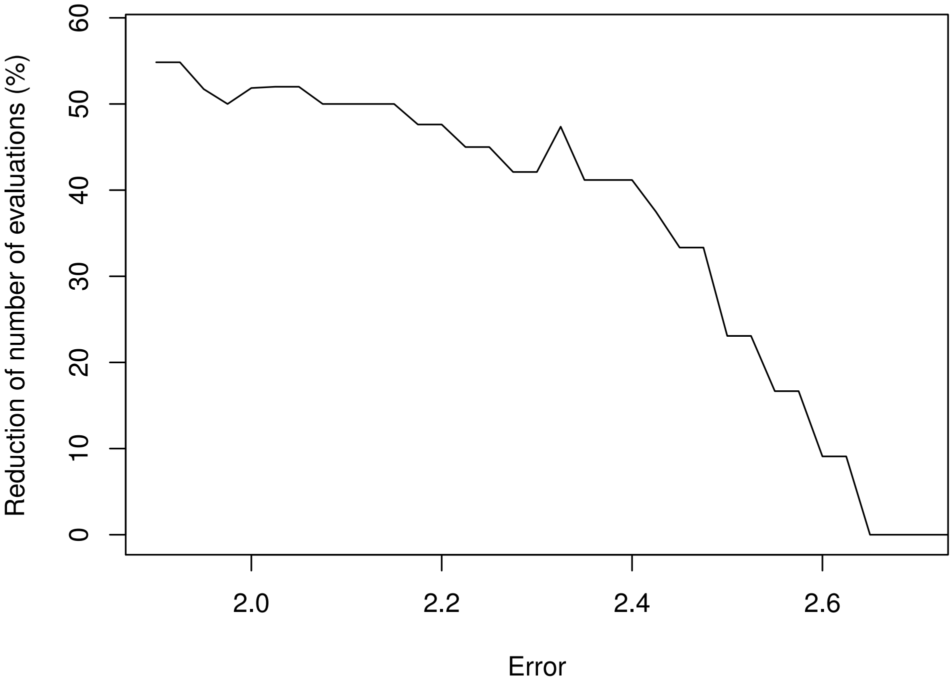


Figure 5.10: Overall estimated warping for the test case 2 (represented by its effect on a regular grid).

up the computation by a factor of about 9.5. As here, speedup with respect to q is typically concave: the more an algorithm is parallelised, the smaller is the gain of incrementing q.

We also note that the speedup given by the non-stationary model is lower than with the stationary one, whereas the non-stationary model still gives lower prediction error (fig. 5.9). Here the non-stationary model is the most effective with small batchsize ($q = 1$ and $q = 5$). A sequential design algorithm with a large batchsize have fewer iterations, and the locations of the evaluations are more dependent on early, less precise models. On the contrary, there are more iterations with a small batchsize, and a more precise model performs even better in the long run as part of the evaluation budget is allocated according to the model.

Another indicator of the performance of multipoint sampling is the "evaluation efficiency" $t_1/(qt_q)$, where t_i is the time required for the i-point sequential algorithm to reach a given precision. The value qt_q can be seen as a measure of the computational resources needed for the q-point method, taking into account the computational time t_q and the number of computers q. Figure 5.13 shows the evaluation efficiency for different precision levels and batchsize. We observe that, when we focus on the computational resource, the efficiency of



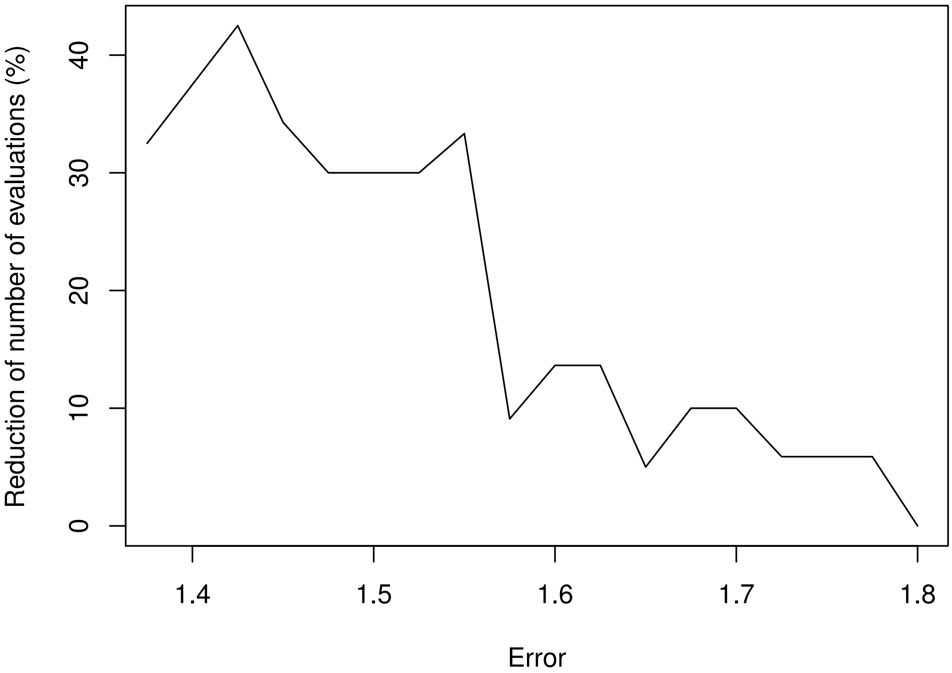


Figure 5.11: Reduction of the number of evaluations gained using a given method: top, reduction with the non-stationary model compared to the stationary model in sequential settings (MSE); bottom, reduction with the sequential multipoint MSE design ($q = 5$) compared to a non-sequential optimised LHS in non-stationary settings. Abscissa is the median error.

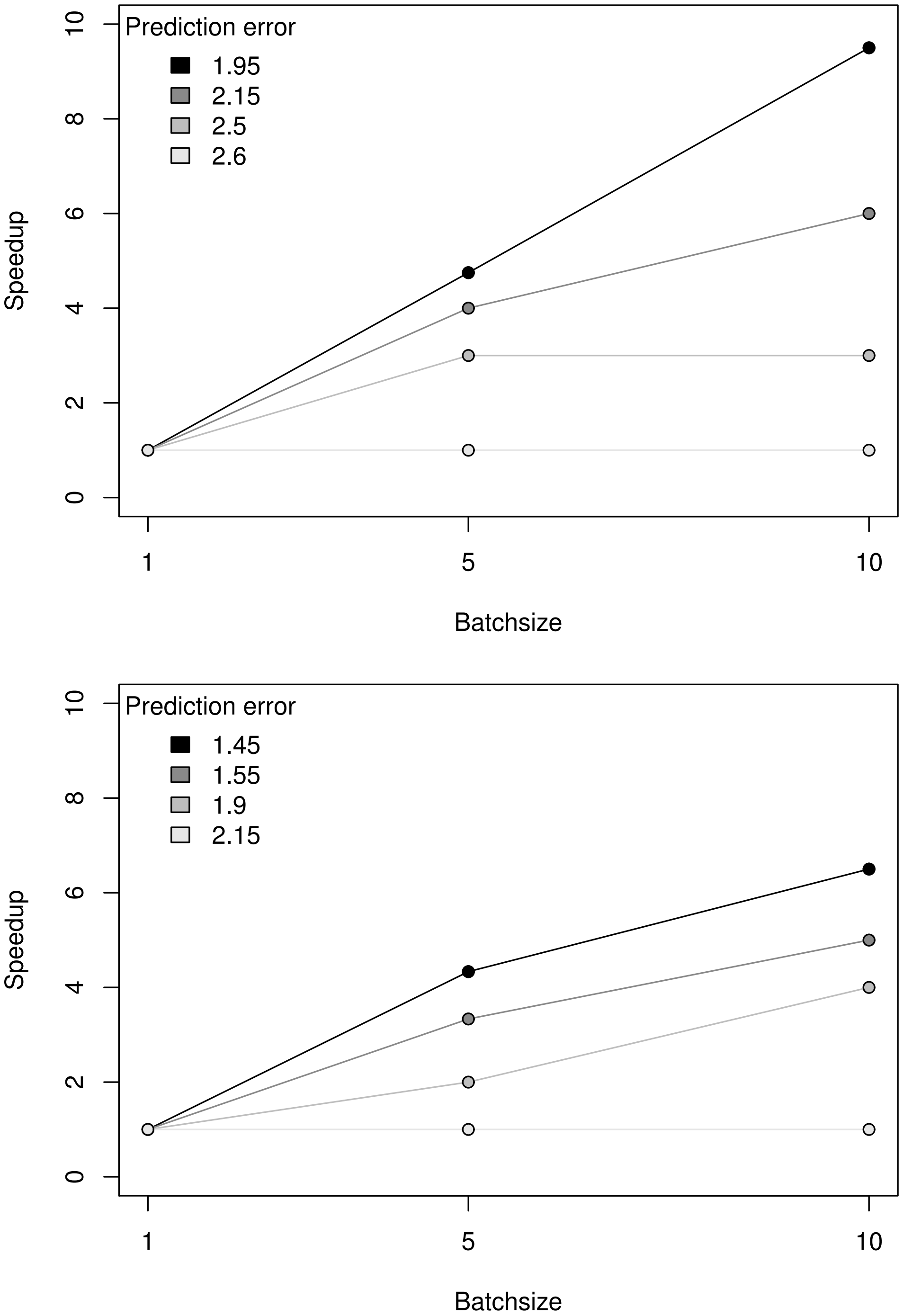


Figure 5.12: Speedup of the sequential design of experiments for the stationary model (top) and the non-stationary (bottom). Calculation of speedup is based on computation times for achieving below a given median prediction error.

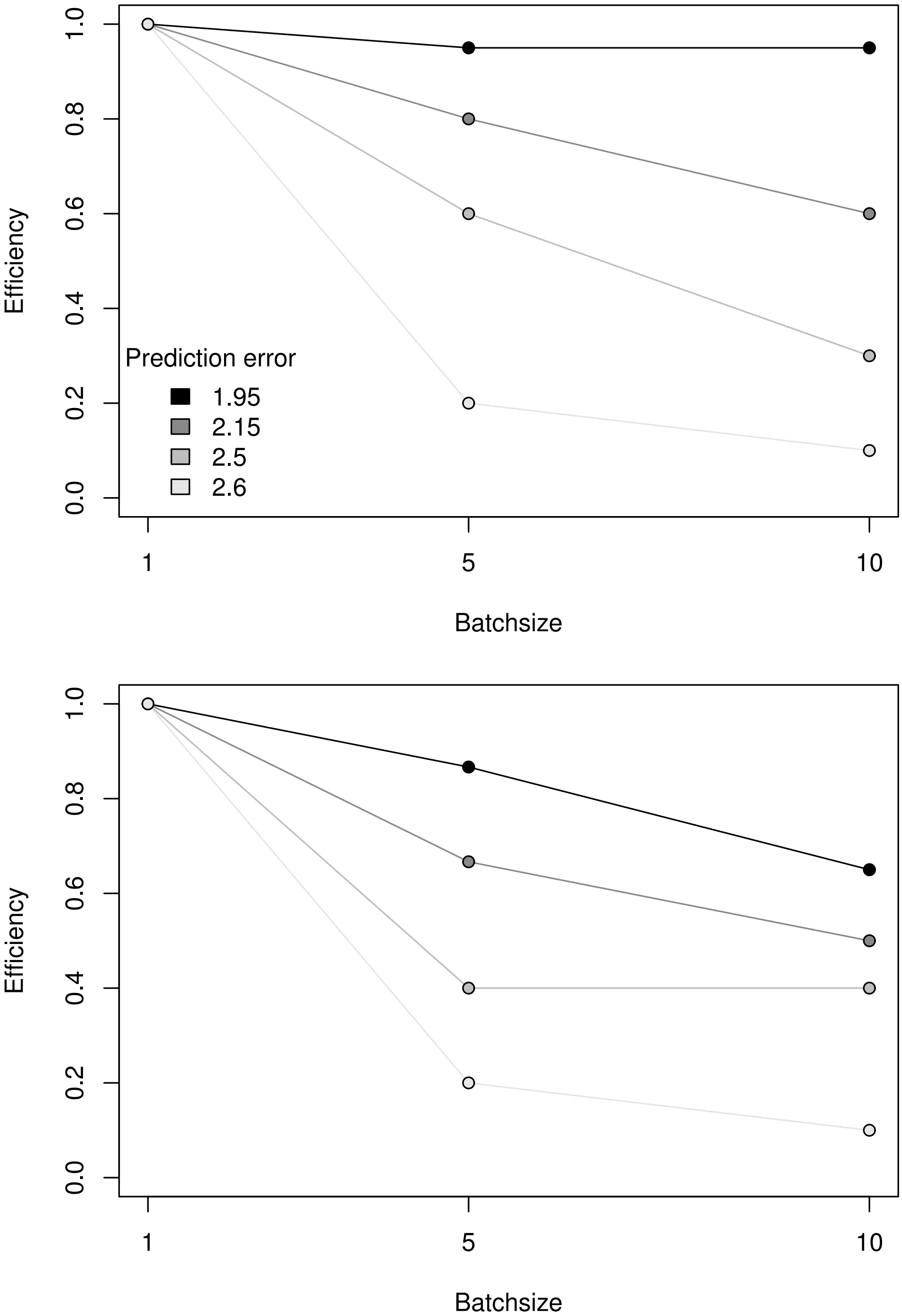


Figure 5.13: Evaluation efficiency of the sequential design of experiments with respect to the batchsize q, for the stationary model (top) and the non-stationary (bottom). Calculation of evaluation efficiency are based on computation times for achieving a given median prediction error.

using a sequential design is lower in the multipoint setting than when $q = 1$ (this is also noticeable in the first panel of fig. 5.9 for $q = 5$). However, in addition to much shorter computation times (see the second panel of fig. 5.9), multipoint sequential settings are shown in fig. 5.11 to nonetheless reduce the prediction errors compared to the non-sequential optimal LHS designs.

5.3.2 Application of the Wav-GP

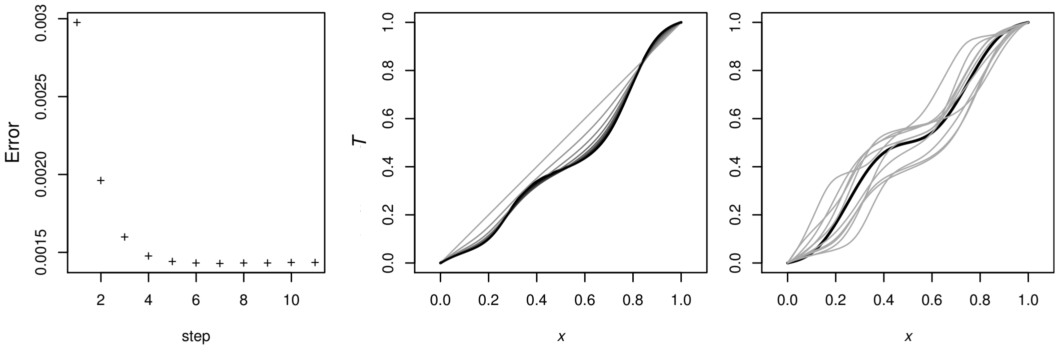


Figure 5.14: Warping estimation. From left to right: 1) error step by step between the warping approximation and the real warping run for a given warped function ; 2) progression step by step (from grey to black) of the estimated warping with the algorithm run for a given warped function; 3) the true warping (black) and 9 warping estimations (grey) resulting from different functions (sample paths of a Matérn kernel, $\nu = 5/2$) warped by the same true warping.

Empirical convergence of the warping estimation algorithm

We show in fig. 5.14 the different iterates of the warping estimation algorithm (see 3.4) as well as the empirical convergence of the L^2 error from a known deformation of a 1D signal on a 10-point design (the underlying stationary signal is a GP realisation such as plotted in the right panel of fig. 5.15 and the deformation is displayed as a black curve in the central panel of fig. 5.14). It turns out that the warping approximation error stabilises quite fast (see the left panel of fig. 5.14, where the error stabilises from the fifth iteration, up to minor variations presumably due to Monte Carlo fluctuations). However, while the error appears to reduce and stabilise, it does not converge to zero, meaning that the algorithm does not lead to the true warping. This was to be expected as we are here attempting to recover a warping map from only

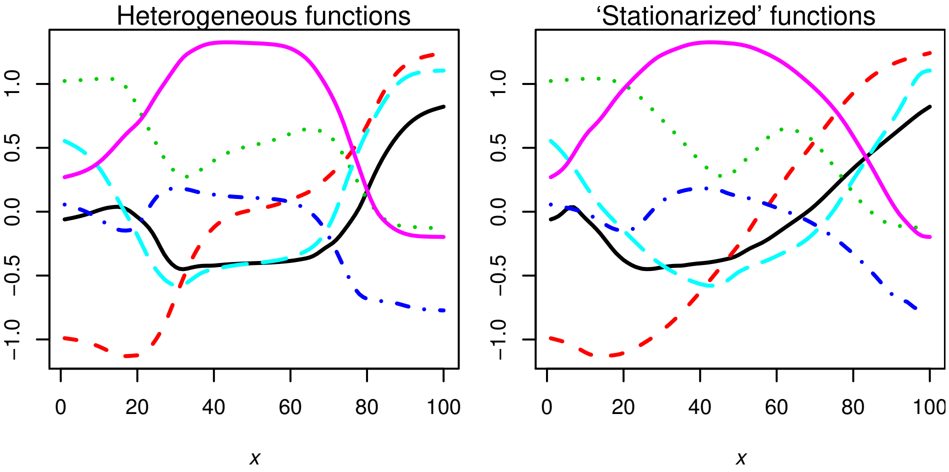


Figure 5.15: Comparison between a real function and its estimated "stationarisation" after five iterations of the algorithm: 1) functions with heterogeneous variations, assumed to be warped by γ, i.e. $f \circ \gamma$; 2) based on 10 evaluation points of the warped functions $f \circ \gamma$, the warping γ^* is estimated and the functions are "stationarised", i.e. $f \circ \gamma \circ \gamma^{*-1}$ is displayed.

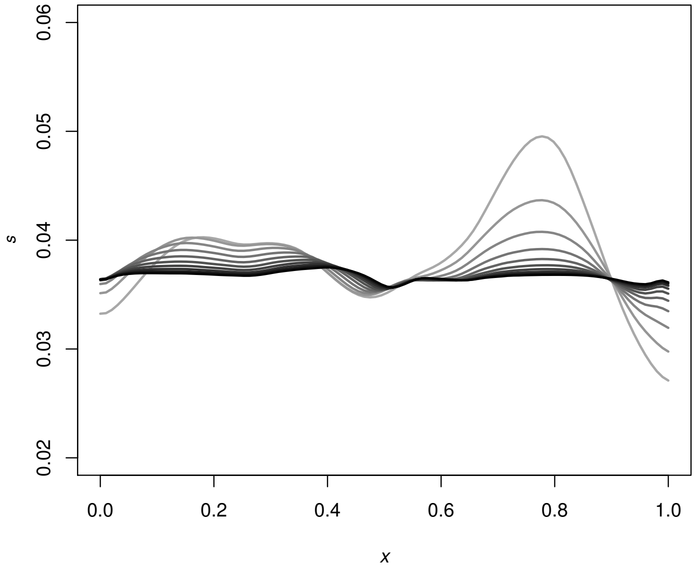


Figure 5.16: Empirical convergence of the local scale ($G_{Y_i}(\cdot)$, section 2.2.4) towards a more constant function during 6 iterations of Algorithm 3.4.

10 observations of the warped function. It would be interesting however to study the behaviour of the reconstruction error curves (after stabilisation of the algorithm) under the GP distribution assumed for the starting signal. While this opens perspectives for future work (be it based on theoretical analysis or on extensive stochastic simulation benchmarking), here we rather focused on applying the approach to a few GP realisations for illustrative purposes and then to focus on real data, as exposed below. In brief, preliminary numerical experiments based on GP simulations suggested that our proposed algorithm offers promising results for stationarizing warped functions relying on scattered evaluations. Following up on fig. 5.14, fig. 5.16 illustrates how the local scale evolves over iterations. If we focus on the function derived after the algorithm has stabilised, the associated local scale is close to constant, corresponding to a mitigation of the warping-induced heterogeneous variations illustrated in fig. 5.15.

Prediction from fixed designs

We test now the performance of the proposed approach on a unidimensional cut (when $\ln(W_2) = 1.3 \times 10^{-3}$) of the IRSN test case 2 data (see fig. 5.17).

The parameters of the estimation algorithm are $N_{\text{stop}} = 4$ and $p = 50$. The proposed approach is compared to a prediction under a stationary GP model, under a warped GP model with parametric warping map expressed as an increasing continuous piecewise second degree polynomial, under a Composite GP model [Ba et al., 2012] and under TGP. The models are run on space filling designs of experiments of size $n = 8$ or $n = 40$. The whole workflow is replicated 50 times with different designs. For each method and design size, a median error and a standard deviation are computed (see Table 5.3). In this test case, as expected the stationary model is the least accurate. Axial warping model have mitigate results. CGP and TGP and our model leads to smaller prediction errors while keeping a moderate standard deviation. Although Wav-GP is here the most accurate, the difference between the medians of CGP, TGP and Wav-GP are still small given the level of the standard deviations[4].

[4] To give a particular perspective, applying Welch's unequal variance t-test, to test if the 50 errors from the Wav-GP model have a lower mean than from TGP or CGP models (with assumption of normality, see [Welch, 1947]), does not show statistical significance, except for Wav-GP over TGP when $n = 40$ (p-value around 3×10^{-4}).

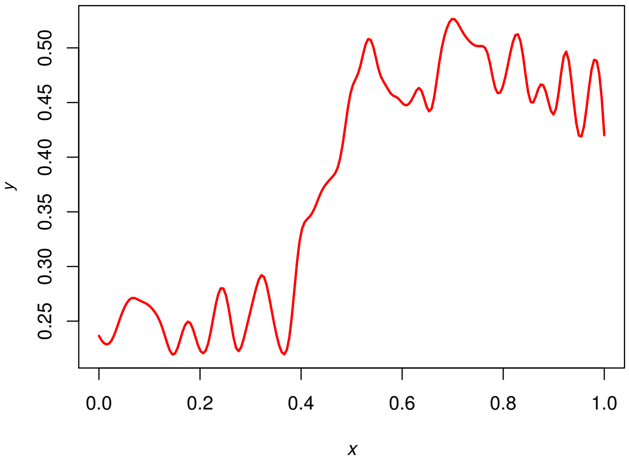


Figure 5.17: Unidimensional cut ($\ln(W_2) = 1.3 \times 10^{-3}$) of the IRSN test case 2.

Table 5.3: Median error and standard deviation for different method and design sizes (fig. 5.17).

	Stat. GP	Warp GP	CGP	TGP	Wav-GP
$n = 8$	3.70 (0.61)	3.50 (0.68)	3.33 (0.36)	3.35 (0.53)	3.30 (0.52)
$n = 40$	1.21 (0.10)	1.17 (0.14)	1.11 (0.19)	1.19 (0.13)	1.09 (0.15)

Multivariate extension

We define and apply here a heuristic extension of the univariate algorithm to the case of a bivariate function whose sections run along an unknown direction and are supposed to present almost the same heterogeneity, calling for an identical univariate warping. First we build a model using a GP Z conditioned on $\mathcal{A}_n$ with a standard covariance, say stationary anisotropic Matérn with smoothness parameter $\nu = 5/2$. Then let us denote by $\boldsymbol{u}$, $||\boldsymbol{u}|| = 1$, the direction in input space along which the function has its (main) heterogeneous variations. While several approaches could be envisaged to estimate $\boldsymbol{u}$ when unprescribed, as here, in this test case it is estimated using the normed eigenvector corresponding to the largest eigenvalue of the positive definite matrix of anisotropy estimated within an initially fitted GP model with a geometric anisotropic covariance (e.g. [Rasmussen and Williams, 2006], see section 2.1.1 about the parametrisation).

We perform a transformation of the input space in the direction $\boldsymbol{u}$:

$$\boldsymbol{\gamma}(\boldsymbol{x}) = P_{\boldsymbol{u},\boldsymbol{x}_0}(\boldsymbol{x}) + \gamma(\langle \boldsymbol{x} - \boldsymbol{x}_0, \boldsymbol{u}\rangle)\boldsymbol{u} \tag{5.4}$$

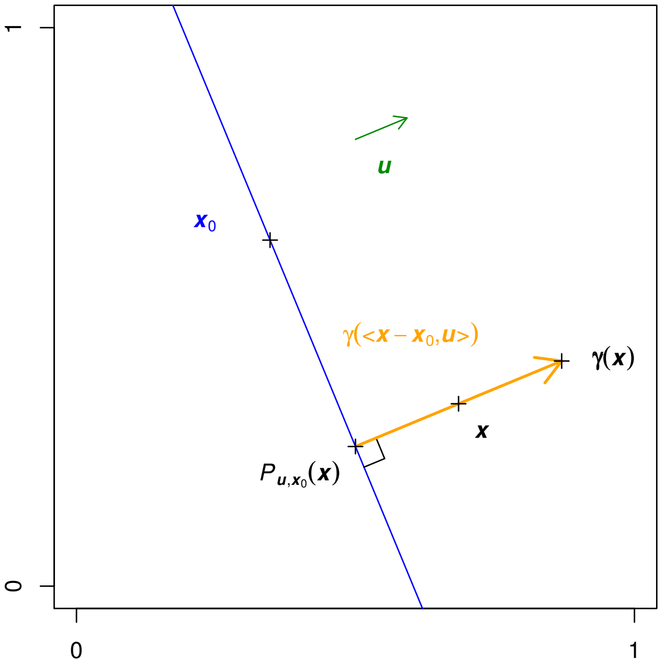
with $P_{\boldsymbol{u},\boldsymbol{x}_0}(\boldsymbol{x}) = \boldsymbol{x} - \langle \boldsymbol{x} - \boldsymbol{x}_0, \boldsymbol{u}\rangle\boldsymbol{u}$ the projection of $\boldsymbol{x}$ on the hyperplane $H_{\boldsymbol{u},\boldsymbol{x}_0}$ with normal vector $\boldsymbol{u}$ and containing a fixed arbitrary point $\boldsymbol{x}_0$ in the considered two-dimensional domain (see fig. 5.18).

The univariate warping $\gamma(\cdot)$ is estimated with the algorithm of section 3.4. As an input of the algorithm, here the initial univariate $Y^{(0)}$ is empirically chosen as random sections of Z conditioned on $\mathcal{A}_n$:

$$Y^{(0)} : t \longrightarrow Z\left(\mathrm{V} + t\boldsymbol{u}\right) \tag{5.5}$$

with V a random vector following a uniform distribution on a closed subset $\mathcal{V}$ of the hyperplane $H_{\boldsymbol{u},\boldsymbol{x}_0}$. We choose $\mathcal{V}$ such that the input domain belongs to $\{\boldsymbol{v} + t\boldsymbol{u}, \boldsymbol{v} \in \mathcal{V}, t \in \mathbb{R}\}$. This operation means that the realisations of $Y^{(0)}$ are univariate sections of Z conditioned on $\mathcal{A}_n$ at random locations and in the direction $\boldsymbol{u}$. Here $Y^{(0)}$ is treated as an input of Algorithm 3.4. The algorithm returns an estimation of the warping γ and the overall multivariate warping of Z is given in eq. (5.4).

The multivariate extension is applied to the IRSN test case 1 and compared with the stationary and CGP models. The parameters of the estimation algorithm are $N_{\text{stop}} = 4$ and $p = 50$. Following the same numerical approach, we build for each method 10000 models from random LHS designs of 20 points. Prediction errors are displayed in fig. 5.19. We observe in terms of median



1
0
0
1
$\boldsymbol{u}$
$\boldsymbol{x}_0$
$\gamma(<\boldsymbol{x}-\boldsymbol{x}_0,\boldsymbol{u}>)$
$\gamma(\boldsymbol{x})$
$\boldsymbol{x}$
$P_{\boldsymbol{u},\boldsymbol{x}_0}(\boldsymbol{x})$

Figure 5.18: Sketch of the transformation of the input space in the direction $\boldsymbol{u}$ according to eq. (5.4).

error (and also for the two other quartile errors), the Wav-GP model is the most efficient one on this test case. Despite an overall best performance, the approach based on Wav-GP is less accurate than other models when focussing on high quantiles. This can be the motivation of further investigation on the robustness of the heuristic extension of the proposed approach.

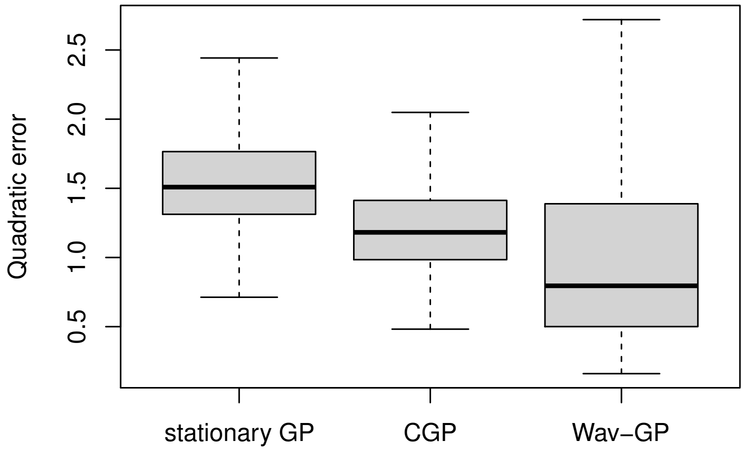


Figure 5.19: Prediction error resulting from three probabilistic models (median, first and third quartiles, and for the 'whiskers', the highest (resp. lowest) value whose absolute difference with the third (resp. the first) quartile is less than 2/3 of the inter quartile distance).

5.4 Global optimisation with gradient of multipoint expected improvement

The goal of this section is to illustrate the usability of the proposed gradient-based q-EI maximisation schemes and in particular the improvements brought by the fast formulas detailed in the previous sections. The relevance of using sequential sampling strategies based on the q-EI maximisation has already been investigated before (see [Chevalier and Ginsbourger, 2014., Wang et al., 2015, Marmin et al., 2015]) and all these articles pointed out the importance of calculation speed which often limits the use of q-EI based strategies to moderate q. We do not aim again at proving the performance of q-EI based sequential strategies. Instead we aim at illustrating the gain, in computation time, brought by the fast formulas and show that using the approximate gradient obtained in eq. (4.31) does not impair the ability to find batches with (close to) maximal q-EI.

Experimental setup: sequential minimisation strategies

We now perform a total of 50 minimisations of f, each using an initial design of experiments of $n_0 = 80$ points drawn from an optimum-LHS procedure with a different seed. Three different batch-sequential strategies are investigated.

The first one – serving as a benchmark – is a variation of the "Constant Liar Mix" heuristic [Chevalier and Ginsbourger, 2014., Wang et al., 2015] where, at each iteration, the batch of size q is chosen among several batches obtained from the Constant Liar heuristic [Ginsbourger et al., 2010] with different lie levels. We use 7 lie levels fixed to the current maximum observation the current minimum observation, and the $2.5\%, 10\%, 50\%, 90\%, 97.5\%$ quantiles of the conditional distribution of the point selected in the batch. A total of 7 batches are proposed at each iteration and the CL-mix heuristic picks the one with maximum q-EI.

The two other strategies considered here rely on pure q-EI maximisation using a multistart BFGS algorithm with a stopping criterion of precision 2.2×10^{-7} (parameter `control$factr` of the R function "optim" [R Core Team, 2015]). The gradients involved in the optimisation are computed either with the tangent moment formula or the proxy. For the gradient-based q-EI maximisation, we use a total of 10 starting batches obtained, again, using a Constant Liar heuristic with random lies sampled from the conditional distribution at the selected point. Finally we use two different batch sizes. When $q = 8$ we run a total of 10 iterations and when $q = 4$ we run 20 iterations. The hyper-parameters of the GP model are re-estimated at each iteration after having incorporated the new observations.

First q-EI maximisation

We first compare the performance, in terms of q-EI, of the multistart BFGS algorithm when the proxy gradient and the tangent moment methods are used. Table 5.4 compares the results at iteration 1 for these two methods and the CL-mix strategy. The results are averaged over the 50 initial designs.

As expected, the CL-mix heuristic yields batches with lower q-EI than the strategies directly maximizing q-EI. Also, for both $q = 4$ and $q = 8$, the

Table 5.4: Average q-EI value of the optimal batches found for each of the 50 initial designs. The numbers between brackets are the average computation times.

	$q = 4$	$q = 8$
tangent moment	12.45 (22.6 s)	15.35 (700.2 s)
proxy	12.46 (14.3 s)	15.35 (127.0 s)
CL-mix	11.80 (7.7 s)	14.34 (15.6 s)

two q-EI based methods have the same performance, which stresses out the relevance of the proxy method since the latter is about 1.6 times faster when $q = 4$ and 5.5 faster when $q = 8$.

Several *q*-EI maximisation steps

We now compare the performances of the different q-EI maximisation approaches after multiple batch evaluations. Figure 5.20 displays the average regret as a function of the iteration number (first row) and the total computation time (i.e. the time to evaluate f and find the next batch to evaluate) assuming respectively that the computation time of f is 0 seconds (i.e. instantaneous), two minutes and one hour (rows $2, 3, 4$ respectively). Looking at the performances as a function of the iteration number (first row on fig. 5.20), the CL-mix heuristic, which samples a batch with lower q-EI at each step, leads in average to a slower convergence than the two other methods, for both $q = 4$ and $q = 8$. In contrast, the two strategies based on q-EI maximisation have similar performances.

However, these conclusions do not hold when the regret is plotted as a function of the total computation time (rows $2, 3, 4$ on fig. 5.20). First, when the computation time t_{eval} of f is null (row 2) it is clear that q-EI-based sequential strategies are not adapted since they are too expensive. In this case, the CL-mix heuristic performs better and some other optimisation strategies which are not metamodel-based would probably be more relevant. Second, when f is moderately expensive (i.e. $t_{\text{eval}} = 2$ minutes), the proxy method and CL-mix have comparable performances when $q = 8$, but the proxy outperforms when $q = 4$. Besides, the proxy shows a much faster convergence than the tangent moment method when $q = 8$. The use of q-EI based strategies thus becomes relevant when t_{eval} is larger than a few minutes, if the proxy is used. Finally, when t_{eval} is equal to one hour, the use of q-EI based strategies is particularly

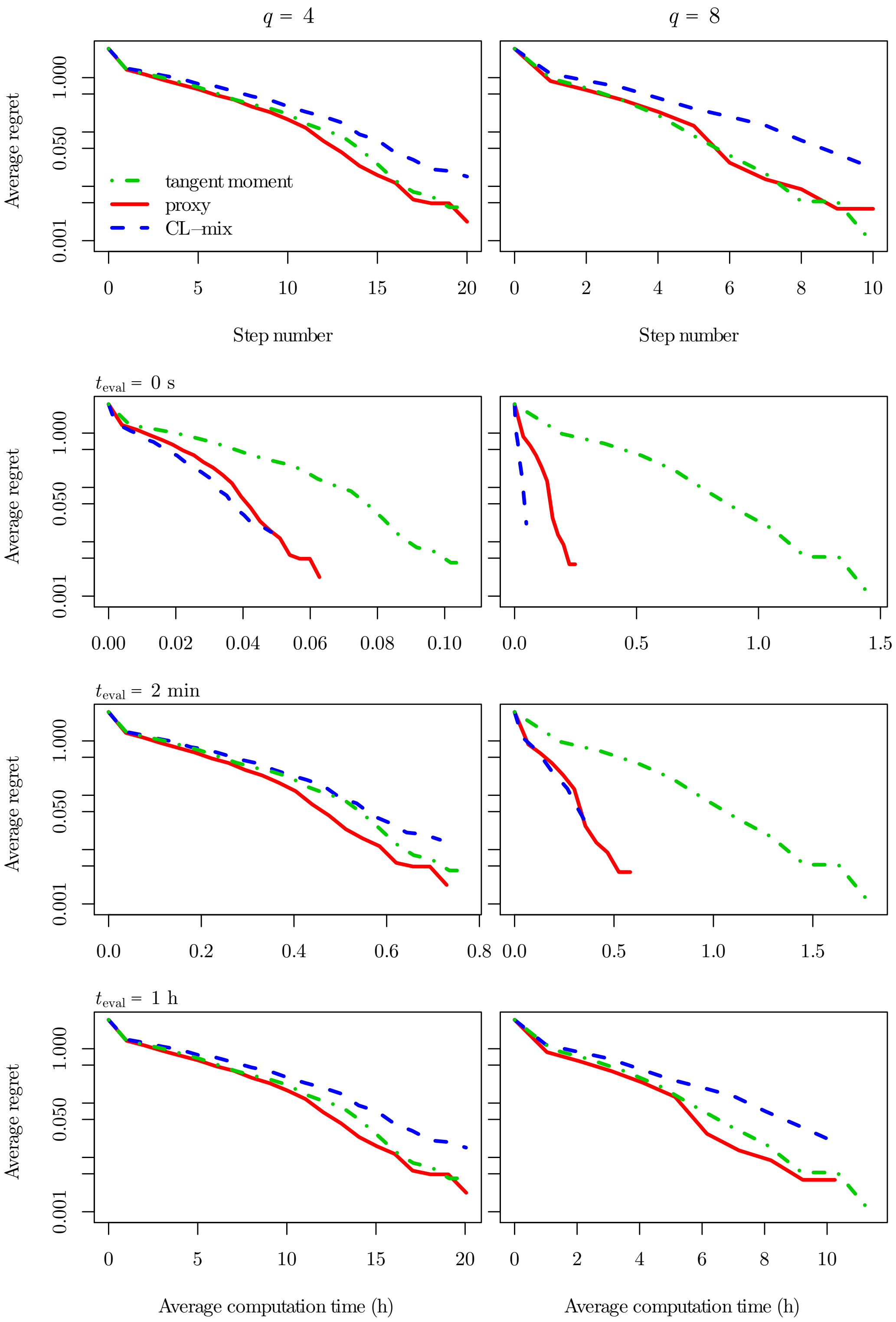


Figure 5.20: Log-scaled average regret of the three considered optimisation strategies as a function of the iteration number (row 1) and the total computation time (rows $2, 3, 4$) assuming that the computation times of f, t_{eval}, are respectively 0 seconds, 2 minutes and 1 hour. Experiments are performed with $q = 4$ (left column) and $q = 8$ (right column).

recommended. In that case the relative improvement of the proxy compared to the tangent moment method tends to naturally vanish because of the long computation time of f. When f is extremely expensive to compute, using the proxy is thus not essential. However, since it does not impair the ability to find a batch with large q-EI we still recommend to use it, especially when q is large.

Conclusion

Motivated by engineering problems, in particular regarding computational costs in some IRSN case studies, we have developed non-stationary modelling and sampling approaches for predicting and designing experiments in a context of function with heterogeneous variations and expensive evaluations.

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 covariance 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 maximisation. It is also relevant to point that in case of a high variation zone slightly misaligned with a canonical axis, our model is favoured in numerical tests because its linear component can estimate an appropriate rotation of the data before the non-linear warping. 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”.

We also have introduced in this thesis wav-GP, an original coupling between wavelets and warped GP models for approximation of function with heterogeneous variations under scattered evaluations. The key ingredient is the estimation of the warping map by combining the computation of the local scale and conditional simulations. An algorithm consisting of successive local scale estimations and stationarisation steps via estimated warping maps was proposed for the univariate case and then extended towards a bivariate application. Numerical tests conducted so-far for $d = 1$ on Gaussian process realisations and on the considered univariate test case have highlighted a fast stabilisation of warping estimates, and also competitive prediction performance compared to several state-of-the-art GP prediction methods on the mechanical test case. The current proposal of bivariate extension, that boils down to revisiting the univariate algorithm along random lines with a chosen direction, leads to competitive median performance on our considered test-case, but at the price of an increased variability compared to the other methods. These first experimental results on the proposed algorithm and its extension call for additional research like a more general multivariate extension. More generally, results obtained here in terms of prediction performance and the highlighted links between local scale and warping are encouraging to further investigate theoretical and methodological connections between local analysis and non-stationary Gaussian process modelling.

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.

Finally, we provide a closed-form expression of the generalised q-points Expected Improvement criterion for batch-sequential Bayesian global optimisation. An interpretation based on moments of truncated Gaussian vectors yields fast q-EI formulas with arbitrary precision. Furthermore a new approximation for the gradient is shown to be even faster while preserving ability to find batches close to maximal q-EI. As the use of these strategies was previously considered cumbersome from a dozen of batch points, these formulas happen to be of particular interest to run q-EI based batch-sequential strategies for larger batch sizes. We show that these methods are implementable and efficient on a classic 8-dimensional test case. Additionally, some of the intermediate results established here might be of interest for other research questions involving moments of truncated Gaussian vectors and their gradients. Perspectives include deriving second order derivatives of q-EI and fast numerical estimates thereof. Also, we aim at improving the sampling of initial batches in multistart derivative-based q-EI maximisation.

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 or weighted IMSE approaches. In the same flavour of using the gradient of the modelling GP, the curvature or the wavelet coefficients of the GP could be used as quantities for deriving new sampling criteria. 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 maximisation, and also model selection versus full Bayesian approaches for inferring the number of lines of the matrix A and further parameters, pertaining for instance to the univariate deformations.

Appendix A

On the multipoint expected improvement for GP-based optimisation

Generalities on the EI criterion

Definition of expected improvement

In recent years, optimisation based on a GP model has attracted a lot of interest from both the computer experiments and machine learning communities. We recall that a GP Y is used for taking into account prior information on f through a trend function $m : \mathcal{D} \to \mathbb{R}$ and a covariance kernel $c : (\boldsymbol{x}, \boldsymbol{x}') : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$. Once m and c are specified, possibly up to some parameters to be inferred based on data (see 2.1), the considered GP model can be used as an instrument to locate the next evaluation point(s) via a criteria. While a number of Bayesian optimisation criteria have been proposed in the literature (see, section 2.1.3 for a short review and e.g., [Jones, 2001, Frazier et al., 2008, Villemonteix et al., 2009, Srinivas et al., 2010, Contal et al., 2014] and references therein), we concentrate here on the *Expected Improvement* (EI) criterion [Mockus, 1989, Jones et al., 1998] and on variations thereof, with a focus on its use in batch-sequential optimisation. Denoting by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{D}$ points where f is assumed evaluated and by $\boldsymbol{x}_{n+1:n+q} := (\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q}) \in \mathcal{D}^q$ a batch of candidate points where to evaluate f next, the multipoint EI of batchsize q (or for short q-EI) is defined as

$$\mathrm{EI}_{n,q}(\boldsymbol{x}_{n+1:n+q}) = \mathbb{E}_n\left(\left(\min_{i=1,\dots,n} Y_{\boldsymbol{x}_i} - \min_{j=n+1,\dots,n+q} Y_{\boldsymbol{x}_j}\right)_+\right), \tag{A.1}$$

where $\mathbb{E}_n$ refers to the conditional expectation knowing the event $\mathcal{A}_n := \{Y_{\boldsymbol{x}_1} = f(\boldsymbol{x}_1), \dots, Y_{\boldsymbol{x}_n} = f(\boldsymbol{x}_n)\}$. One way of calculating such criterion is to rely on Monte Carlo simulations. Figure A.1 illustrates both what the criterion means and how to approach it by simulations, relying on three samples from the multivariate Gaussian distribution underlying eq. (A.1).

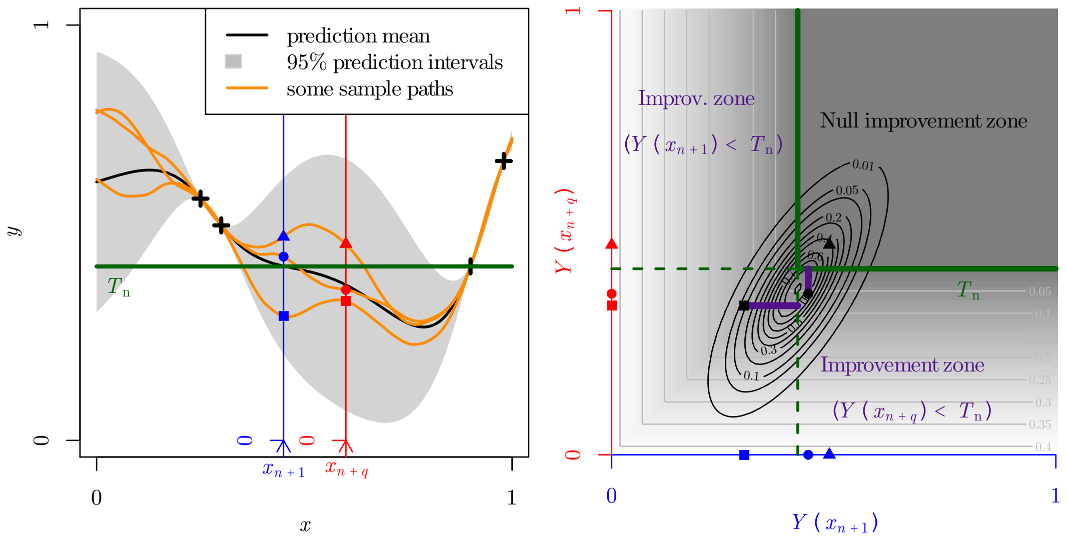


Figure A.1: Illustration of the principles underlying q-EI for $d = 1$, $n = 4$, $q = 2$. Left: Gaussian process prediction of a function f from observations $\mathcal{A}_n$ (depicted by black crosses). The horizontal line stands for $\boldsymbol{\gamma}_n$, the smallest response value from $\mathcal{A}_n$. Three conditional simulation draws are plotted in light orange and various point symbols represent their respective values at two unobserved locations x_{n+1} and x_{n+q}. Right: distribution of the random vector $\left(Y_{x_{n+1}}, Y_{x_{n+q}}\right)^\top$ knowing $\mathcal{A}_n$ (black contours). For each point symbol, the length of the thick purple segment represents the improvement realised by the corresponding sample path. The multipoint EI is the expectation of this length, or in other words, it is the integral of the improvement (grey-scale function) with respect to the conditional distribution of $\left(Y_{x_{n+1}}, Y_{x_{n+q}}\right)^\top$ knowing $\mathcal{A}_n$.

Analytical derivation of expected improvement

Now, for $q = 1$, it is well known that EI can be expressed in closed form as a function of the posterior mean and variance m_n and $\sigma_n : \boldsymbol{x} \to c_n(\boldsymbol{x}, \boldsymbol{x})$ as follows

$$\mathrm{EI}_{n,1}(\boldsymbol{x}) = \begin{cases} u_n(\boldsymbol{x})\Phi(u_n(\boldsymbol{x})) \quad +\sigma_n(\boldsymbol{x})\left(\varphi(u_n(\boldsymbol{x})/\sigma_n(\boldsymbol{x}))\right) & \text{if } \sigma_n(\boldsymbol{x}) \neq 0 \\ \quad (u_n(\boldsymbol{x}))_+ & \text{otherwise} \end{cases} \tag{A.2}$$

where $u_n(\boldsymbol{x}) = \min_{i=1,\dots,n} f(\boldsymbol{x}_i) - m_n(\boldsymbol{x})$ and Φ, φ are the cumulative distribution function and probability density function of the standard Gaussian distribution, respectively.

When deriving eq. (A.2) (hence for $q = 1$), eq. (A.1) happens to involve a first order moment of the truncated univariate Gaussian distribution. As shown in [Chevalier and Ginsbourger, 2014.], it turns out that eq. (A.1) can be expanded in a similar way in the multipoint case ($q \geqslant 2$) relying on moments of truncated Gaussian vectors.

The applied motivation for having batch-sequential EI algorithms is strong, as distributing evaluations of Bayesian optimisation algorithms over several computing units allows significantly reducing wall-clock time and with the fast popularization of clouds, clusters and GPUs in recent years it is becoming always more commonplace to launch several calculations in parallel. Even at a slightly inflated price and scripting effort, reducing the total time is often a primary goal in order to deliver conclusions involving heavy experiments, be they numerical or laboratory experiments, in studies subject to hard time limitations. Obviously, given its practical importance, the question of parallelizing EI algorithms and alike by selecting $q > 1$ points per iteration has been already tackled in a number of works from various disciplinary horizons (including notably [Queipo et al., 2006, Azimi et al., 2010, Desautels et al., 2012, Contal et al., 2013, González et al.]). In this thesis we essentially focus on approaches relying on the maximization of eq. (A.1) and related multipoint criteria, notably by deriving closed-form formulas and fast approximates in section 4.2. On this topic, the multipoint EI of eq. (A.1) has been first calculated in closed form for the case $q = 2$ in [Ginsbourger et al., 2010]. For the case $q \geqslant 3$, a Monte Carlo scheme and some sub-optimal batch selection strategies were proposed. Further work on Monte Carlo simulations for multipoint EI estimation can be found in [Janusevskis et al., 2012, Girdziusas et al., 2012]; besides this, stochastic simulation ideas have been explored in [Frazier, 2012] for maximizing this multipoint EI criterion via a stochastic gradient algorithm, an approach recently investigated in [Wang et al., 2015]. Meanwhile, a closed-form formula for the multipoint EI relying on combinations of $(q-1)$-and q-dimensional Gaussian cumulative distribution functions was obtained in [Chevalier and Ginsbourger, 2014.], a formula which applicability in reasonable time is however restricted to moderate q (say $q \leqslant 10$) in the current situation. Building upon [Chevalier and Ginsbourger, 2014.], [Marmin et al., 2015] recently calculated the gradient of the multipoint EI criterion in closed form and obtained some first experimental results on (non-stochastic) gradient-based multipoint EI maximization.

Calculations for multipoint EI and its gradient

Differentiating multivariate Gaussian CDF

We consider the CDF dimension $p \geqslant 2$. We use the convention $\Phi_0 = 1$.

Gradient

Using the following identity, derived from conditional distributions of a Gaussian vector,

$$\forall i = 1, \dots, p, \ \varphi_{p,\Sigma}\left(\boldsymbol{x}\right) = \varphi_{1,\Sigma_{ii}}\left(x_i\right) \varphi_{p-1,\Sigma_{|_i}}\left(\boldsymbol{x}_{-i} - \boldsymbol{m}_{|i,x_i}\right),$$

with $\boldsymbol{m}_{|i,u} = \frac{u}{\Sigma_{ii}} \boldsymbol{\Sigma}_{-i,i}$ and $\Sigma_{|i} = \Sigma_{-i,-i} - \frac{1}{\Sigma_{ii}} \boldsymbol{\Sigma}_{-i,i} \boldsymbol{\Sigma}_{-i,i}^{\top}$, we reformulate the integral of the Gaussian CDF:

$$\forall i = 1, \dots, p, \Phi_{p,\Sigma}\left(\boldsymbol{x}\right) = \int_{-\infty}^{x_i} \varphi_{1,\Sigma_{ii}}\left(u_i\right) \Phi_{p-1,\Sigma_{|i}}\left(\boldsymbol{x}_{-i} - \boldsymbol{m}_{|i,u_i}\right) \mathrm{d}u_i.$$

Here indexed minus symbols, e.g. in $\boldsymbol{\Sigma}_{-i,i}$, refer to exclusions of a line or a column.

Finally we have

$$\nabla \Phi_{p,\Sigma}\left(\boldsymbol{x}\right) = \left(\varphi_{1,\Sigma_{ii}}\left(x_i\right) \Phi_{p-1,\Sigma|i}\left(\boldsymbol{x}_{-i} - \boldsymbol{m}_{|i,x_i}\right)\right)_{i=1,\dots,p}. \tag{A.3}$$

Hessian

As for the computation of the gradient, we write
$\forall i, j = 1, \ldots, p, i \neq j$,

$$\Phi_{p,\Sigma}\left(\boldsymbol{x}\right) = \int\limits_{-\infty}^{x_i}\int\limits_{-\infty}^{x_j} \varphi_{2,\Sigma_{ij,ij}}\left(\begin{bmatrix}u_i\\u_j\end{bmatrix}\right)\Phi_{p-2,\Sigma_{|ij}}\left(\boldsymbol{x}_{-\{i,j\}}-\boldsymbol{m}_{|(i,j),(u_i,u_j)}\right)\mathrm{d}u_j\mathrm{d}u_i,$$

with $\boldsymbol{m}_{|(i,j),(u,u')} = \Sigma_{-\{ij\},ij}\Sigma_{ij,ij}^{-1}\begin{bmatrix}u\\u'\end{bmatrix}$ and $\Sigma_{|ij} = \Sigma_{-\{ij\},-\{ij\}}-\Sigma_{-\{ij\},ij}\Sigma_{ij,ij}^{-1}\Sigma_{-\{ij\},ij}^{\top}$.

So $\forall i, j = 1, \ldots, p, i \neq j$,

$$\frac{\partial^2\Phi_q}{\partial x_i \partial x_j}\left(\boldsymbol{x}\right) = \varphi_{2,\Sigma_{ij,ij}}\left(\begin{bmatrix}x_i\\x_j\end{bmatrix}\right)\ \Phi_{p-2,\Sigma_{|ij}}(\boldsymbol{x}_{-\{i,j\}}-\boldsymbol{m}_{|(i,j),(x_i,x_j)}).$$

When $i = j$, the differentiation of eq. (A.3) gives,

$$\frac{\partial^2\Phi_q}{\partial x_i^2}\left(\boldsymbol{x}\right) = -\frac{1}{\Sigma_{ii}}\left(x_i\frac{\partial\Phi_{p,\Sigma}}{\partial x_i}(\boldsymbol{x}) + \sum_{\substack{j=1\\j\neq i}}^{p}\Sigma_{ij}\frac{\partial^2\Phi_{p,\Sigma}}{\partial x_i \partial x_j}\left(\boldsymbol{x}\right)\right).$$

Moments of truncated multivariate Gaussian distribution

Analytical formula (propositions 10 and 15)

We see here why we can derive an analytical formula of $\mathcal{M}_{r,\alpha}(\boldsymbol{m},\Sigma)$, with $r \leqslant s \in \mathbb{N}\backslash\{0\}$, $\boldsymbol{m} \in \mathbb{R}^s$ and $\Sigma \in S_{++}^s$, by differentiating $\mathcal{G}$, defined in eq. (4.18). It is known, see e.g. Cressie et al. [1981], that moments can be obtained differentiating the moment generating function $G_{\boldsymbol{m},\Sigma,s}$:

$$\mathcal{M}_{r,\alpha}(\boldsymbol{m},\Sigma) = \frac{\partial^\alpha G_{\boldsymbol{m},\Sigma,s}}{\partial t_r^\alpha}(\mathbf{0}),$$

with, for $r \in \{1,\ldots,s\}$, $G_{\boldsymbol{m},\Sigma,r} : \boldsymbol{t} \to \mathbb{E}\left(\exp\left(\boldsymbol{t}^\top\boldsymbol{Z}\right)\mathbb{1}_{(Z_1,\ldots,Z_r)^\top\leqslant\mathbf{0}}\right)$, $\boldsymbol{Z} \sim \mathcal{N}\left(\boldsymbol{m},\Sigma\right)$. We derive now an analytical formula for $G_{\boldsymbol{m},\Sigma,r}$. As needed in proposition 15, we derive an analytical formula for any r, and not only for $r = s$.

$\forall \boldsymbol{t} \in \mathbb{R}^s$,

$$\begin{aligned}
G_{\boldsymbol{m},\Sigma,r}(\boldsymbol{t}) &= \overbrace{\int_{-\infty}^{0}\!\!\cdots\!\int_{-\infty}^{0}}^{r \text{ times}} \int_{-\infty}^{\infty}\!\!\cdots\!\int_{-\infty}^{\infty} \exp\left(\boldsymbol{t}^\top \boldsymbol{z}\right)\varphi_\Sigma\left(\boldsymbol{z} - \boldsymbol{m}\right)\ \mathrm{d}z_1 \dots \mathrm{d}z_s \\
&= \varphi_\Sigma(\boldsymbol{0}) \int_{-\infty}^{0}\!\!\cdots\!\int_{-\infty}^{0} \int_{-\infty}^{\infty}\!\!\cdots\!\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\left(\boldsymbol{z} - \boldsymbol{m}\right)^\top \Sigma^{-1}\left(\boldsymbol{z} - \boldsymbol{m}\right) - 2\boldsymbol{t}^\top \boldsymbol{z}\right)\right)\mathrm{d}\boldsymbol{z} \\
&= e^{-\frac{1}{2}\left(-\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)^\top \Sigma\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)+\boldsymbol{m}^\top\Sigma^{-1}\boldsymbol{m}\right)} \\
&\qquad \varphi_\Sigma(\boldsymbol{0}) \int_{-\infty}^{0}\!\!\cdots\!\int_{-\infty}^{0} \int_{-\infty}^{\infty}\!\!\cdots\!\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left((\boldsymbol{z}-\boldsymbol{m}-\Sigma\boldsymbol{t})^\top\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{m}-\Sigma\boldsymbol{t})^\top\right)}\mathrm{d}\boldsymbol{z} \\
&= e^{\frac{1}{2}\left(\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)^\top \Sigma\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)-\boldsymbol{m}^\top\Sigma^{-1}\boldsymbol{m}\right)} \Phi_{r,\left(\Sigma_{ij}\right)_{i,j\leqslant r}}\left(-\boldsymbol{m} - \left(\Sigma_{ij}\right)_{i\leqslant r, j\leqslant s}\boldsymbol{t}\right).
\end{aligned} \tag{A.4}$$

In the frame of the proof of proposition 15,

- if $\Sigma_r(\boldsymbol{u},\boldsymbol{v})$, the covariance matrix of $(\boldsymbol{Z}_{\boldsymbol{v}}^\top, Z_r(\boldsymbol{u}))^\top$, is positive definite, we take
$$M_{\boldsymbol{u},\boldsymbol{v}} = t \to G_{(\boldsymbol{m}(\boldsymbol{v}), m_r(\boldsymbol{u})),\Sigma_r(\boldsymbol{u},\boldsymbol{v}),p}((0,\dots,0,t)^\top),$$
- else, as $\Sigma(\boldsymbol{v})$ is definite positive, there exists only one index r_0 such as $Z_r(\boldsymbol{u}) = Z_{r_0}(\boldsymbol{v})$ almost surely (for example $r = r_0$ when $\boldsymbol{u} = \boldsymbol{v}$), and we have
$$M_{\boldsymbol{u},\boldsymbol{v}} = t \to G_{(\boldsymbol{m}(\boldsymbol{v})),\Sigma(\boldsymbol{v}),p}((0,\dots,\underset{\substack{\uparrow \\ r_0^{\text{th}} \text{ position}}}{t},\dots,0)^\top).$$

  Equation (A.4) leads to eq. (4.30) in both cases.

Differentiation with respect to mean and covariance

We differentiate here eq. (4.20) with respect to $\boldsymbol{m}$ and Σ.

With respect to the mean $\boldsymbol{m}$

$$\frac{\partial \mathcal{M}_{r,1}}{\partial \boldsymbol{m}}(\boldsymbol{m},\Sigma) = \Phi_{p,\Sigma}(-\boldsymbol{m})\boldsymbol{e}_r - m_r\nabla\Phi_{p,\Sigma}(-\boldsymbol{m}) + \nabla\nabla^\top\Phi_{p,\Sigma}(-\boldsymbol{m})\boldsymbol{\Sigma}_r. \tag{A.5}$$

With respect to the covariance Σ

$$\frac{\partial \mathcal{M}_{r,1}}{\partial \Sigma}(\boldsymbol{m},\Sigma) = m_r \frac{\partial}{\partial \Sigma}\Phi_{p,\Sigma}(-\boldsymbol{m}) - \sum_{i=1}^{p} \varphi_{\Sigma_{ii}}\left(-m_i\right)\Phi_{p-1,\Sigma|i}(-\boldsymbol{m}_{|i})E^{(r,i)}$$
$$+\Sigma_{ri}\frac{\partial}{\partial \Sigma_{ii}}\varphi_{\Sigma_{ii}}\left(-m_i\right)\Phi_{p-1,\Sigma|i}(-\boldsymbol{m}_{|i})E^{(i,i)}$$
$$+\Sigma_{ri}\varphi_{\Sigma_{ii}}\left(-m_i\right)\frac{\partial}{\partial \Sigma}\Phi_{p-1,\Sigma|i}(-\boldsymbol{m}_{|i}). \tag{A.6}$$

with $\boldsymbol{m}|_i = \boldsymbol{m}_{-i} - \frac{m_i}{\Sigma_{ii}}\boldsymbol{\Sigma}_{-i,i}$ and $\Sigma|_i = \Sigma_{-i,-i} - \frac{1}{\Sigma_{ii}}\boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}^{\top}_{-i,i}$. Writting $\mathrm{d}_\Sigma\left[\boldsymbol{m}|_i\right]$ $\mathrm{d}_\Sigma\left[\Sigma|_i\right]$ the differential of the functions $\Sigma \to \boldsymbol{m}|_i$ and $\Sigma \to \Sigma|_i$, we have:

$$\mathrm{d}_\Sigma\left[\boldsymbol{m}|_i\right](H) = \frac{m_i}{\Sigma_{ii}}\boldsymbol{H}_{-i,i} \tag{A.7}$$

$$\mathrm{d}_\Sigma\left[\Sigma|_i\right](H) = H_{-i,-i} + \frac{H_{ii}}{\Sigma_{ii}^2}\boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}^{\top}_{-i,i} - \frac{2}{\Sigma_{ii}}\boldsymbol{H}_{-i,i}\boldsymbol{\Sigma}^{\top}_{-i,i} \tag{A.8}$$

$$\frac{\partial}{\partial \Sigma}\Phi_{p-1,\Sigma|i}(-\boldsymbol{m}_{|i}) = \sum_{r=1}^{p}\sum_{s=1}^{p}\Big(-\mathrm{d}_\Sigma\left[\boldsymbol{m}|_i\right](E^{(r,s)}).\nabla\Phi_{p-1,\Sigma|_i}(-\boldsymbol{m}|_i)$$
$$+\mathrm{tr}\left(\frac{\partial}{\partial \Gamma}\Phi_{p-1,\Sigma|i}\left(-\boldsymbol{m}_{|i}\right).\mathrm{d}_\Sigma\left[\Sigma_{|i}\right]\left(E^{(r,s)}\right)\right)\Big)E^{(r,s)}$$

with:

- $E^{(r,s)} = (\delta_{ij})_{i,j=1,\dots,p}$,
- $\frac{\partial}{\partial \Gamma}\Phi_{p-1,\Sigma_{|i}}(-\boldsymbol{m}_{|i})$ the derivative of $\Gamma \to \Phi_{p-1,\Gamma}(-\boldsymbol{m}_{|i})$ evaluated at $\Sigma_{|i}$. We use the Plackett's differential equation, extended by Berman [1987], to find
$$\frac{\partial}{\partial \Gamma}\Phi_{p-1,\Sigma_{|i}}(-\boldsymbol{m}_{|i}) = \nabla\nabla^{\top}\Phi_{p-1,\Sigma_{|i}}(-\boldsymbol{m}_{|i}),$$
$\nabla\nabla^{\top}\Phi$ is given in appendix A.

Generalized *q*-EI as a sum of moments

Proof. For given (ℓ, r) in $\{1,\dots,n\}\times\{1,\dots,q\}$, we consider $E_{\ell,r}$ the event that the random variable inside the expectation term of eq. (4.16) equals $\left(Y_{\boldsymbol{x}_\ell} - Y_{\boldsymbol{x}_{n+r}}\right)^{\alpha}$. We have

$$E_{\ell,r} = \{Y_{\boldsymbol{x}_{n+r}} \leqslant Y_{\boldsymbol{x}_\ell}\} \;\cap\; \{\forall i \leqslant n, i \neq \ell; Y_{\boldsymbol{x}_\ell} \leqslant Y_{\boldsymbol{x}_i}\}$$
$$\cap\; \left\{\forall j \leqslant q, j \neq r; Y_{\boldsymbol{x}_{n+r}} \leqslant Y_{\boldsymbol{x}_{n+j}}\right\}$$

Considering all pairs $(\ell,\, r)$, we have:

$$EI_n(\boldsymbol{x}_{n+1:n+q}) = \sum_{\ell=1}^{n}\sum_{r=1}^{q} \mathbb{E}_n\left(\left(Y_{\boldsymbol{x}_\ell} - Y_{\boldsymbol{x}_{n+r}}\right)^\alpha \mathbb{1}_{E_{\ell,r}}\right).$$

For each term (ℓ, r) of the sum, the conditioning event can be rewritten $E_{\ell,r} = \{\ \boldsymbol{Z}^{(\ell,r)}(\boldsymbol{x}_{n+1:n+q}) \leqslant \boldsymbol{0}\}$, with $\boldsymbol{Z}^{(\ell,r)}$ a random vector of size $n+q-1$, defined by the following linear transformation of $\boldsymbol{Y} = \left(Y_{\boldsymbol{x}_1}, \dots, Y_{\boldsymbol{x}_{n+q}}\right)^\top$:

$$\forall i = 1, \dots, n{+}q{-}1,\, Z_i^{(\ell,r)} = \begin{cases} Y_\ell - Y_i & \text{if } 1 \leqslant i \leqslant \ell - 1 \\ Y_\ell - Y_{i+1} & \text{if } \ell \leqslant i \leqslant n-1 \\ Y_{n+r} - Y_{i+1} & \text{if } n \leqslant i \leqslant n+q-1, i \neq n+r-1 \\ Y_{n+r} - Y_\ell & \text{if } i = n+r-1 \end{cases}$$

Indeed, the first $n{-}1$ components of $\boldsymbol{Z}^{(l,r)} \leqslant \boldsymbol{0}$ reflect $\{\forall i \leqslant n, i \neq \ell; Y_{\boldsymbol{x}_\ell} \leqslant Y_{\boldsymbol{x}_i}\}$, and the last components reflect $\left\{\forall j \leqslant q, j \neq r; Y_{\boldsymbol{x}_{n+r}} \leqslant Y_{\boldsymbol{x}_{n+j}}\right\}$ and $\{Y_{\boldsymbol{x}_{n+r}} \leqslant Y_{\boldsymbol{x}_\ell}\}$.

□

Mean square differentiability of $\boldsymbol{Y_x^\alpha \mathbb{1}_B}$

Let B be an event, Y be a mean-squared differentiable Gaussian process and $\alpha \in \mathbb{N}$. Then we have:

$$\mathbb{E}\left(\left(\frac{Y_{x+h}^\alpha - Y_x^\alpha}{h}\mathbb{1}_B - \frac{\mathrm{d}Y^\alpha}{\mathrm{d}x}(x)\mathbb{1}_B\right)^2\right)$$
$$\leqslant \mathbb{E}\left(\left(\frac{Y_{x+h}^\alpha - Y_x^\alpha}{h} - \frac{\mathrm{d}Y^\alpha}{\mathrm{d}x}(x)\right)^2\right) \underset{h\to 0}{\longrightarrow} 0$$

by mean-squared differentiability of Y^α.

Symmetry argument

The term $\frac{q(q+1)}{2}$ comes from a symmetry occurring when summing terms with different index but actually equal. At fixed summation index ℓ in eq. (4.22), we denote ω_{ri} the i^{th} term in the scalar product in eq. (4.20) for each $\mathcal{M}_{m+r-1,1}$ required for q-EI:

$$\forall i, r = 1, \dots, q,\ \omega_{ri} = \Sigma^{(\ell,r)}_{ri} \left[\nabla \Phi_{p,\Sigma^{(\ell,r)}}(-\boldsymbol{m}^{(\ell,r)})\right]_i .$$

Then the following symmetry between indices i and r occurs:

$$\forall i, r = 1, \dots, q,\ \frac{\omega_{ri}}{\Sigma^{(\ell,r)}_{ri} \varphi_{1,\Sigma^{(\ell,r)}_{ii}}(-m^{(\ell,r)}_i)} = \frac{\omega_{ir}}{\Sigma^{(\ell,i)}_{ir} \varphi_{1,\Sigma^{(\ell,i)}_{rr}}(-m^{(\ell,i)}_r))}$$

Indeed, using the formula of the derivative of CDF, (appendix A), leads to:

$$\begin{aligned}\frac{\omega_{ri}}{\Sigma^{(\ell,r)}_{ri} \varphi_{\Sigma^{(\ell,r)}_{ii}}(-m^{(\ell,r)}_i))} &= \Phi_{p-1,\Sigma^{(\ell,r)}_{|i}}(-\boldsymbol{m}^{(\ell,r)}_{|i}) \\ &= \mathbb{P}\left(\begin{array}{c|c} Y_{\boldsymbol{x}_{n+r}} \leqslant Y_{\boldsymbol{x}_\ell}, & Y_{\boldsymbol{x}_{n+i}} = \\ Y_{\boldsymbol{x}_{n+j}} \leqslant Y_{\boldsymbol{x}_{n+r}}, \forall j = 1 \dots q, j \neq r, j \neq i & Y_{\boldsymbol{x}_{n+r}} \end{array}\right),\end{aligned}$$

which is clearly symmetrical between i and r.

Appendix B

Short introduction on Stepwise Uncertainty Reduction (SUR)

We focus on Stepwise Uncertainty Reduction (SUR) strategies. Its aim is to provide an optimal sequence of evaluation points, selecting each point in order to reduce a uncertainty quantity. This approach requires a precise definition of a uncertainty function H_n. The function $H_n : (\mathcal{D} \times \mathbb{R})^n \to \mathbb{R}^+$ gives the remaining uncertainty after n evaluations $y_i = Y_{\boldsymbol{x}_i}$, with $(\boldsymbol{x}_i, y_i)_{i=1,\dots,n} \in (\mathcal{D} \times \mathbb{R})^n$. Thus, the main difference between SUR methods is the definition of the uncertainty H_n that often relies on a GP model.

The SUR strategy focus on the remaining uncertainty at depletion of the available budget of N points. Because the design of experiments cannot be known before evaluating the random process Y, the evaluations points $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ are considered as realisations of random vectors[1] $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ in $\mathcal{D}$. In an optimal SUR framework, the realisations $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ as well as the realisations $(y_1, \dots, y_N)$ are obtained sequentially, with for all $n = 0, \dots, N-1$,

$$\boldsymbol{x}_{n+1} \in \underset{\boldsymbol{x} \in \mathcal{D}}{\operatorname{argmin}} \ \mathbb{E}\left(H_N\left(\left(\boldsymbol{X}_1, Y_{\boldsymbol{X}_1}\right), \dots, \left(\boldsymbol{X}_N, Y_{\boldsymbol{X}_N}\right)\right) \middle| \mathcal{A}_n \cup \{\boldsymbol{X}_{n+1} = \boldsymbol{x}\}\right). \tag{B.1}$$

with $\mathcal{A}_n$ the event $\{(\boldsymbol{X}_1, Y_{\boldsymbol{X}_1}) = (\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{X}_n, Y_{\boldsymbol{X}_n}) = (\boldsymbol{x}_n, y_n)\}$, $\mathcal{A}_0 = \varnothing$. This approach is numerically complex, as the computation of this expectation at a single $\boldsymbol{x}$ involves the distributions of $\boldsymbol{X}_i$, minimisers of entangled stochastic processes.

[1] The $\boldsymbol{X}_i$'s are in capital to emphasise that they are random and in bold font to emphasise that they are vectors (and not matrices of batches of points).

Some papers propose practical designs of experiments using SUR, see e.g. [Bect et al., 2011, González et al., 2016]. The one-step-lookahead simplification is the most straightforward way to get tractable sampling criteria. With an appropriate choice of uncertainty function H_{n+1}, it encompasses previously cited sampling criteria as, among others, (multipoint) EI [Ginsbourger and Le Riche, 2010] and IMSE. The idea is to select the next evaluation by minimising the expected uncertainty at the next step, i.e. by replacing N by $n+1$ in eq. (B.1),

$$\boldsymbol{x}_{n+1} \in \underset{\boldsymbol{x}\in\mathcal{D}}{\operatorname{argmin}}\ \mathbb{E}\left(H_{n+1}\left(\left(\boldsymbol{x}_1, y_1\right), \ldots, \left(\boldsymbol{x}_n, y_n\right), \left(\boldsymbol{x}, Y_{\boldsymbol{x}}\right)\right)\right). \tag{B.2}$$

More generally for multipoint sampling, we define

$$\begin{aligned} &X_{n+1:n+q} \in \underset{\breve{x}_{n+1},\ldots,\breve{x}_{n+q}\in\mathcal{D}}{\operatorname{argmin}} \\ &\mathbb{E}\left(H_{n+q}\left(\left(\boldsymbol{x}_1, y_1\right), \ldots, \left(\boldsymbol{x}_n, y_n\right), \left(\breve{\boldsymbol{x}}_{n+1}, Y_{\breve{\boldsymbol{x}}_{n+1}}\right), \ldots, \left(\breve{\boldsymbol{x}}_{n+q}, Y_{\breve{\boldsymbol{x}}_{n+q}}\right)\right)\right). \end{aligned} \tag{B.3}$$

For more details on SUR strategies see e.g. [Bect et al., 2016].

Bibliography

R. J. Adler. *The geometry of random fields*, volume 62. Siam, 2010.

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

S. Arlot. V-fold cross-validation improved: V-fold penalization. *arXiv preprint arXiv:0802.0566*, 2008.

J. Azimi, A. Fern, and X. Fern. Batch bayesian optimization via simulation matching. In *Advances in Neural Information Processing Systems*, 2010.

A. Azzalini and A. Genz. *The R package `mnormt`: The multivariate normal and t distributions (version 1.5-1)*, 2014.

D. Azzimonti. *Contributions to Bayesian set estimation relying on random field priors*. PhD thesis, University of Bern, 2016.

S. Ba and V. R. Joseph. *CGP: Composite Gaussian process models*, 2014. R package.

S. Ba, V. R. Joseph, et al. Composite gaussian process models for emulating expensive functions. *Annals of Applied Statistics*, 6(4):1838–1860, 2012.

R. Battiti and F. Masulli. Bfgs optimization for faster and automated supervised learning. In *International neural network conference*, pages 757–760. Springer, 1990.

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.

J. Bect, F. Bachoc, and D. Ginsbourger. A supermartingale approach to gaussian process based sequential design of experiments. *arXiv preprint arXiv:1608.01118*, 2016.

J. Bect, L. Li, and E. Vazquez. Bayesian subset simulation. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):762–786, 2017.

S. M. Berman. An extension of Plackett's differential equation for the multivariate normal density. *SIAM Journal on Algebraic Discrete Methods*, 8(2): 196–197, 1987.

B. J. Bichon, M. S. Eldred, L. P. Swiler, S. Mahadevan, and J. M. McFarland. Efficient global reliability analysis for nonlinear implicit performance functions. *American Institute of Aeronautics and Astronautics Journal*, 46(10): 2459–2468, 2008.

M. Binois, D. Ginsbourger, and O. Roustant. Quantifying uncertainty on pareto fronts with gaussian process conditional simulations. *European Journal of Operational Research*, 243(2):386–394, 2015.

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

Roberto Calandra, Jan Peters, Carl Edward Rasmussen, and Marc Peter Deisenroth. Manifold gaussian processes for regression. In *2016 International Joint Conference on Neural Networks (IJCNN)*, page 3338–3345. IEEE, 2016. doi: 10.1109/IJCNN.2016.7727626. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7727626.

C. Chevalier and D. Ginsbourger. *Learning and Intelligent Optimization - 7th International Conference, Lion 7, Catania, Italy, January 7-11, 2013, Revised Selected Papers*, chapter fast computation of the multipoint expected improvement with applications in batch selection, pages 59-69. *Springer*, 2014.

C. Chevalier, D. Ginsbourger, J. Bect, and I. Molchanov. Estimating and quantifying uncertainties on level sets using the vorob?ev expectation and deviation with gaussian process models. In *mODa 10–Advances in Model-Oriented Design and Analysis*, pages 35–43. Springer, 2013.

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, 2014a.

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, 2014b.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

T. Choi, J. Q. Shi, and B. Wang. A gaussian process regression approach to a single-index model. *J. Nonparametr. Stat.*, 23(1):21–36, 2011.

M. Clerc and S. Mallat. Estimating deformations of stationary processes. *Annals of Statistics*, 31(6):1772–1821, 2003.

E. Contal, D. Buffoni, A. Robicquet, and N. Vayatis. Parallel gaussian process optimization with upper confidence bound and pure exploration. In *European Conference on Machine Learning*, 2013.

E. Contal, V. Perchet, and N. Vayatis. Gaussian process optimization with mutual information. In *Proceedings of the 31st International Conference on Machine Learning*, pages 253–261, 2014.

N. Cressie, A.S. Davis, and J. Leroy Folks. The moment-generating function and negative integer moments. *The American Statistician*, 35 (3):148–150, 1981.

D. Ginsbourger and V. Picheny and O. Roustant and with contributions by C. Chevalier and S. Marmin and T. Wagner. *DiceOptim: Kriging-Based Optimization for Computer Experiments*, 2015. R package version 1.5.

I. Daubechies. *Ten lectures on wavelets.* SIAM, 1992.

T. Desautels, A. Krause, and J. Burdick. Parallelizing exploration-exploitation trade-offs with Gaussian process bandit optimization. In *Proceedings of ICML*, 2012.

Y. Deville, D. Ginsbourger, and O. Roustant. *Package R 'kergp'*, 2015. Contributor : N. Durrande.

J. Doob. Stochastic processes. *New York: John Wiley & Sons*, 1953.

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 and Data Analysis*, 54(4):858–862, 2010.

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.

N. Durrande, D. Ginsbourger, and O. Roustant. Additive covariance kernels for high-dimensional gaussian process modeling. In *Annales de la Faculté des Sciences de Toulouse*, volume 21, pages 481–499, 2012.

D. Eddelbuettel and C. Sanderson. Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, 2014.

K.-T. Fang, R. Li, and A. Sudjianto. *Design and modeling for computer experiments*. Chapman & Hall / CRC Press, 2006.

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

P. Flandrin. *Temps-fréquence*. Hermes, 1993.

A. I. J. Forrester, A. Sóbester, and A. J. Keane. *Engineering design via surrogate modelling: a practical guide*. Wiley, 2008.

P. I. Frazier. Parallel global optimization using an improved multi-points expected improvement criterion. In *INFORMS Optimization Society Conference, Miami FL*, 2012.

P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–149, 1992.

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

D. Ginsbourger and R. Le Riche. Towards gaussian process-based optimization with finite time horizon. *Advances in Model-Oriented Design and Analysis*, 9(96):89–96, 2010.

D. Ginsbourger, R. Le Riche, and L. Carraro. Kriging is well-suited to parallelize optimization. In *Computational Intelligence in Expensive Optimization Problems*, volume 2 of *Adaptation Learning and Optimization*, pages 131–162. Springer, 2010.

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, 2016a.

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, 2016b.

R. Girdziusas, R. Le Riche, F. Viale, and D. Ginsbourger. Parallel budgeted optimization applied to the design of an air duct. Technical report, 2012.

J. González, Z. Dai, P. Hennig, and N. D. Lawrence. Batch bayesian optimization via local penalization. arXiv:1505.08052.

J. González, M. Osborne, and N. Lawrence. Glasses: Relieving the myopia of bayesian optimisation. In *Artificial Intelligence and Statistics*, pages 790–799, 2016.

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

R. B. Gramacy. tgp: An r package for bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models. *Journal of Statistical Software*, 19(9):1–46, 2007.

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.

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

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

R. B. Gramacy and M. Taddy. Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an r package for treed gaussian process models. *Journal of Statistical Software*, 33(6):1–48, 2010.

C.-A. Guérin. Wavelet analysis and covariance structure of some classes of non-stationary processes. *Journal of Fourier Analysis and Applications*, 6 (4):403–425, 2000.

M. S. Handcock and M. L. Stein. A bayesian analysis of kriging. *Technometrics*, 35(4):403–410, 1993.

W. V. Harper and S. K. Gupta. Sensitivity/uncertainty analysis of a borehole scenario comparing latin hypercube sampling and deterministic sensitivity approaches. *Battelle Memorial Institute, Columbus, USA*, (BMI/ONWI–516), 1983.

C. Helbert, D. Dupuy, and L. Carraro. Assessment of uncertainty in computer experiments from universal to bayesian kriging. *Applied Stochastic Models in Business and Industry*, 25(2):99–113, 2009.

D. Higdon. Space and space-time modeling using process convolutions. *Quantitative methods for current environmental issues*, 3754:37–56, 2002.

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

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

J. Janusevskis, R. Le Riche, D. Ginsbourger, and R. Girdziusas. Expected improvements for the asynchronous parallel global optimization of expensive functions : Potentials and challenges. In *LION 6 Conference (Learning and Intelligent OptimizatioN), Paris : France*, 2012.

D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(21):345–383, 2001.

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.

Q. Y. Kenny, W. Li, and A. Sudjianto. Algorithmic construction of optimal symmetric latin hypercube designs. *Journal of Statistical Planning and Inference*, 90(1):145–159, 2000.

O. Knill. *Probability and stochastic processes with applications.* 1994.

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 and Data Analysis*, 53(4):853–856, 2009.

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

S. Mallat. *A Wavelet Tour of signal processing.* Academic Press, USA, 1998.

S. Marmin, C. Chevalier, and D. Ginsbourger. *Machine Learning, Optimization, and Big Data: First International Workshop, MOD 2015, Taormina, Sicily, Italy, July 21-23, 2015, Revised Selected Papers*, chapter Differentiating the Multipoint Expected Improvement for Optimal Batch Design, pages 37–48. Springer International Publishing, Cham, 2015.

G. Matheron. The intrinsic random functions, and their applications. *Advances in Applied Probability*, 5:439–468, 1973.

J. Mockus. *Bayesian Approach to Global Optimization. Theory and Applications.* Kluwer Academic Publisher, Dordrecht, 1989.

A. O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society, Series B*, 40(1):1–42, 1978.

H. Omer and B. Torresani. Time-frequency and time-scale analysis of deformed stationary processes, with application to non-stationary sound modeling. *Applied and Computational Harmonic Analysis*, 10, 2016.

M. Osborne. *Bayesian Gaussian Processes for Sequential Prediction, Optimization and Quadrature.* PhD thesis, University of Oxford, 2010.

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

C. Paciorek and M. Schervish. Nonstationary covariance functions for gaussian process regression. *Advances in Neural Information Processing Systems*, 16: 273–280, 2004.

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.

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.

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

F. Perales, S. Bourgeois, A. Chrysochoos, and Y. Monerie. Two field multibody method for periodic homogenization. *Engineering Fracture Mechanics*, (75), 2008.

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 Mechanics* , 19:389–417, 2010.

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.

V. Picheny, D. Ginsbourger, Y. Richet, and G. Caplin. Quantile-based optimization of noisy computer experiments with tunable precision. *Technometrics*, 55(1):2–13, 2013.

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

N. V. Queipo, A. Verde, S. Pintos, and R. T. Haftka. Assessing the value of another cycle in surrogate-based optimization. In *11th Multidisciplinary Analysis and Optimization Conference*. AIAA, 2006.

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/.

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

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

S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, 59(4):731–792, 1997.

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. *American Institute of Aeronautics and Astronautics Journal*, 4229, 2003.

O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization. *Journal of Statistical Software*, 51(1):1–55, 2012.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435, 1989.

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

T. J. Santner, B. J. Williams, and W. Notz. *The design and analysis of computer experiments*. Springer Sci. & Bus. Media, 2003.

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

M. Schonlau. *Computer Experiments and global optimization*. PhD thesis, University of Waterloo, 1997.

M. C. Shewry and H. P. Wynn. Maximum entropy sampling. *Journal of Applied Statistics*, 1(2):165–170, 1987.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 2012.

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.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022, 2010.

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

M. A. Taddy, H. K. H. Lee, G. A. Gray, and J. D. Griffin. Bayesian guided pattern search for robust local optimization. *Technometrics*, 51(4):389–401, 2009.

G. M. Tallis. The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society, Series B*, 23(1):223–229, 1961.

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.

J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

J. Wang, S. C. Clark, E. Liu, and P. I. Frazier. Parallel bayesian global optimization of expensive functions. Working paper (http://people.orie.cornell.edu/pfrazier/publications.html), 2015.

B. L. Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

B. A. Worley. Deterministic uncertainty analysis. *Transactions of the American Nuclear Society*, 55(CONF-8711195-), 1987.

A. Wu, M. C. Aoi, and J. W. Pillow. Exploiting gradients and hessians in bayesian optimization and bayesian quadrature. *arXiv preprint arXiv:1704.00060*, 2017.

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

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.