DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization

Olivier Roustant, David Ginsbourger, Yves Deville

DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization

Olivier Roustant
Ecole des Mines de St-Etienne

David Ginsbourger
University of Bern

Yves Deville
Alpestat

Abstract

Two recently released R packages for the approximation and the optimization of expensive-to-evaluate deterministic functions are presented. After a motivation of the **DiceKriging** and **DiceOptim** packages relying on an overview of existing needs and softwares in the field of Kriging for Computer Experiments, a self-contained mini-tutorial on Kriging-based approximation and optimization techniques is proposed. The functionalities of both packages are then detailed and demonstrated in two disctinct sections. In particular, the versatility of **DiceKriging** with respect to trend and noise specifications, covariance parameter estimation, as well as conditional and unconditional simulations are illustrated on the basis of several reproducible numerical experiments. The implementation of sequential and parallel optimization strategies relying on the Expected Improvement criterion are then put to the fore on the occasion of **DiceOptim**'s presentation. An appendix is finally dedicated to complementary mathematical and computational details.

Keywords: Computer Experiments, Gaussian Processes, Global Optimization.

1. Introduction

Numerical simulation has become a standard tool in natural science and engineering. Exploited as cheaper and faster complement to physical experiments, simulations can also sometimes be used as necessary substitute to them, e.g. for investigating the behaviour of mechanical structures, or the extreme risks associated with geological storage (e.g. CO_2 sequestration or nuclear waste deposit). A first step to such quantitative investigations is to proceed to a fine mathematical modeling of the phenomenon under study, and to express the function of interest as solution to a set of equations (generally PDE's). Modern simulation techniques such as finite-elements solvers and Monte-carlo methods can then be used to derive approximate solutions to the latter equations. The price to pay in order to derive accurate simulation results is computation time. Conception studies based on an exhaustive exploration of the input space (say on a fine grid) are thus generally impossible under realistic industrial time

constraints. Parcimonious evaluation strategies are hence required, all the more crucially that the computation times and the dimensionality of inputs are high. Mathematical approximations of the input/output relation —also called *surrogate models* or *metamodels*— are increasingly used as a tool to guide simulator evaluations more effeciently. Once a class of surrogate models has been chosen according to some prior knowledge, they can be built upon available observations, and evolve when new data is assimilated. Typical classes of surrogate models for computer experiments include linear regression, splines, neural nets, and Kriging. Here we essentially focus on several variants of the Kriging metamodel, and on their use in prediction and optimization for costly computer experiments.

Originally coming from geosciences (Krige (1951)) and having become the starting point of geostatistics (Matheron (1963)), Kriging is basically a spatial interpolation method. Assuming that an unknown function $y : D \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ is one sample of a real-valued random process $(Y(\mathbf{x}))_{\mathbf{x}\in D}$ with given or partially estimated probability law, Kriging consists in making predictions of unknown $y(\mathbf{x}^{(0)})$ values ($\mathbf{x}^{(0)} \in D$) based on the conditional law of $Y(\mathbf{x}^{(0)})$ knowing observations of Y at a design of experiments $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ ($n \in \mathbb{N}$). Skipping further technical precisions at this stage, let us mention the remarkable fact that the Kriging predictor is interpolating the observations, provided that they are assumed noise-free. It is undoubtedly one of the reasons why this metamodel has been imported from its geostatistical craddle ($d = 2$ or 3) to the high-dimensional framework of computer experiments ($d \in \mathbb{N}$). Indeed, following the impulse given by the seminal paper Sacks, Welch, Mitchell, and Wynn (1989), many works dedicated to Kriging as a surrogate to computer experiments have been published including Welch, Buck, Sacks, Wynn, Mitchell, and Morris (1992) about screening, Koehler and Owen (1996) with an emphasis on experimental design, or O'Hagan (2006) and numerous achievements in uncertainty assessment by the same author since the nineties. To get a global overview of state-of-the-art results and developments in Kriging for computer experiments, we refer to contemporary books of reference such as Santner, Williams, and Notz (2003), Fang, Li, and Sudjianto (2006), and Rasmussen and Williams (2006).

Since the goal of computer experiments is often to tune some control variables in order to optimize a given output, it is obviously tempting to replace a costly simulator by a Kriging metamodel in order to speed-up optimization procedures. Numerical experiments presented in Jones (2001) show that directly optimizing the Kriging predictor is however generally not efficient, and potentially leading to artifactual bassins of optimum in case of iterated optimizations with metamodel update. Fortunately, efficient criteria like the Expected Improvement (EI) have been proposed for sequential Kriging-based optimization, as discussed and compared with other criteria in the latter article, and in Schonlau (1997) and Sasena, Papalambros, and Goovaerts (2002). Already proposed in previous works (e.g. in Mockus (1988)), EI has become an increasingly popular criterion since the publication of the EGO algorithm in Jones, Schonlau, and Welch (1998). Recent advances concern the extension of EI to a multipoints criterion as well as Kriging-based parallel optimization strategies, such as proposed in Ginsbourger, Le Riche, and Carraro (2010). More detail and some additional perspectives can be found in the recent tutorial by Brochu, Cora, and de Freitas (2009).

Back to the implementation of Kriging, let us give a short overview of existing codes, and motivate the emergence of new open source packages for Computer Experiments. Statistical software dedicated to Kriging have flourished since the 1990's, first in the framework of low-dimensional spatial statistics, and later on in computer experiments and optimization. Several

R packages like **spatial**, **geoR**, and **RandomFields** propose indeed a wide choice of functionalities related to classical 2- and 3-dimensional geostatistics. These packages are unfortunately not suitable for applications in higher dimensions, for which similar Kriging equations but specific parameter estimation techniques have to be used. Alternatively, MatLab toolboxes emanating from the computer experiments community have become popular among practitioners, like **MPerK** and **DACE**, or **GPML** in the context of gaussian process regression and classification for machine learning. More recently, some first **R** packages in this vein have emerged, like **mlegp** or **tgp**. **mlegp** possesses good performances in parameter estimation and applications of Kriging for sensitivity analysis and for stochastic filtering, but is restricted to Kriging with gaussian correlation function and first degree polynomial trend, and does not offer any Kriging-based optimization algorithm. In other respects, **tgp** focuses on *treed gaussian process* models, with a bayesian treatment of covariance parameters, and includes a variant of EGO relying on tgp models. Such highly sophisticated metamodels, relying on Markov Chain Monte Carlo techniques, are quite calcultation intensive. Here we consider simpler *Universal Kriging* (UK) metamodels, and put a particular focus on a quick and efficient parameter estimation. The point is to enable the user to optimally benefit from the Universal Kriging potentialities, e.g. through a versatile syntax for the trend structure, and to push the limit of applicability of UK in higher dimensions thanks to a careful implementation of Likelihood maximization routines. In addition, we specifically aim at providing efficient kriging-based optimizers in the spirit of EGO, with optional features for parallel computing.

DiceKriging and **DiceOptim** have been produced in the frame of the DICE consortium. DICE has joined major french companies and public institutions having a high R&D interest in computer experiments with academic participants during the years 2006-2009. The main aim was to put in common industrial problems and academic know-how to foster research and transfer in the field of design of computer experiments. 4 **R** packages summarizing a substantial part of the conducted research have been released on the CRAN at the end of the consortium. The four packages **DiceDesign**, **DiceKriging**, **DiceEval**, and **DiceOptim** should be seen as a small software suite, tailored for different but complementary needs of computer experimenters. For instance, **DiceDesign** might be a good option to generate some original space-filling designs at an initial stage of metamodel fitting with **DiceKriging**, while **DiceEval** might be used to assess the coherence and the fidelity of the obtained metamodel to the observed data. **DiceOptim** is more a complement of **DiceKriging** dedicated to Expected Improvement functions and related sequential and parallel Kriging-based optimization routines. **DiceDesign** and **DiceEval**, shall be presented by their auhors in separated papers. They will not be used nor detailed here.

The main aim of this article is to give a practical introduction to **DiceKriging** and **DiceOptim**, together with an overview of their statistical background. In order to produce a self-contained but portable document, we have chosen to recall basic assumptions in the body of the document (in particular in section 2), and send the reader to the appendix and the bibliography for more technical contents (implementation notes, optimization details, etc.). An overview of the main functionalities of the packages is given in section 3. Section 4 and 5 are dedicated to illustrated examples with code chunks of **DiceKriging** and **DiceOptim**, respectively.

2. Statistical background

Previous to presenting the **DiceKriging** and **DiceOptim** packages from a user perspective, let

us recall some statistical basics of Kriging metamodeling and Kriging-based optimization.

2.1. From Simple to Universal Kriging for deterministic simulators

In this section, the simulator response y is assumed to be a deterministic real-valued function of the d-dimensional variable $\mathbf{x} = (x_1, \ldots, x_d) \in D \subset \mathbb{R}^d$. y is assumed to be one realization of a square-integrable random process $(Y_\mathbf{x})_{\mathbf{x}\in D}$ with first and second moments known up to some parameters. Let us recall that $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ denote the points where y has already been evaluated, and denote by $\mathbf{y} = (y(\mathbf{x}^{(1)}), \ldots, y(\mathbf{x}^{(n)}))'$ the corresponding outputs. For any $\mathbf{x} \in D$, the aim of Kriging will be to optimally predict $Y_\mathbf{x}$ by a linear combination of the observations $\mathbf{y}$. Simple Kriging and Universal Kriging constitute the best linear predictors in two particular cases concerning the process Y and what is known about it.

Simple Kriging: from spatial linear interpolation to Gaussian Process conditioning

In Simple Kriging (SK), Y is assumed to be the sum of a known deterministic trend function $\mu : \mathbf{x} \in D \longrightarrow \mu(\mathbf{x}) \in \mathbb{R}$ and of a centered square-integrable process Z:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}), \tag{1}$$

where Z's covariance kernel $C : (\mathbf{u}, \mathbf{v}) \in D^2 \longrightarrow C(\mathbf{u}, \mathbf{v}) \in \mathbb{R}$ is known. Most of the time, Z is assumed second order stationary, so that $C(\mathbf{u}, \mathbf{v}) = \sigma^2 R(\mathbf{u} - \mathbf{v}; \boldsymbol{\psi})$ where the so-called correlation function R is a function of positive type with parameters $\boldsymbol{\psi}$, and σ^2 is a scale parameter called the process variance. More concerning these parameters can be found in a forthcoming subsection. Concerning Y's trend, note that it is very easy to come back to the centered process $Y(\mathbf{x}) - \mu(\mathbf{x})$ since μ is known. Without loss of generality, we will hence first assume that Y is centered. Let us now derive the best linear unbiased predictor of $Y(\mathbf{x})$ based on the observations $Y(\mathbf{X})$. In other words, let us find $\boldsymbol{\lambda}^*(\mathbf{x}) \in \mathbb{R}^n$ minimizing the mean squared error $\mathrm{MSE}(\mathbf{x}) := \mathbb{E}\left[\left(Y(\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x})' Y(\mathbf{X})\right)^2\right]$. Convexity of MSE ensures both existence and uniqueness of $\boldsymbol{\lambda}^*(\mathbf{x})$, and the SK weights can then be obtained: $\boldsymbol{\lambda}^*(\mathbf{x}) = \mathbf{C}^{-1}\mathbf{c}(\mathbf{x})$, where $\mathbf{C} = (C(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}))_{1\leq i,j\leq n}$ is the covariance matrix of $Y(\mathbf{X})$, and $\mathbf{c}(\mathbf{x}) = (C(\mathbf{x}, \mathbf{x}^{(i)}))_{1\leq i\leq n}$ is the vector of covariances between $Y(\mathbf{x})$ and $Y(\mathbf{X})$. Substituting both the random vector $Y(\mathbf{X})$ by its realization $\mathbf{y}$ and $\boldsymbol{\lambda}(\mathbf{x})$ by $\boldsymbol{\lambda}^*(\mathbf{x})$ in the expression $\boldsymbol{\lambda}(\mathbf{x})' Y(\mathbf{X})$, we get the so-called SK mean prediction at $\mathbf{x}$: $m_{SK}(\mathbf{x}) := \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}\mathbf{y}$. Similarly, by plugging in the optimal $\boldsymbol{\lambda}^*(\mathbf{x})$ in the expression of the MSE, one gets the so-called SK variance at $\mathbf{x}$: $s^2_{SK}(\mathbf{x}) := C(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}\mathbf{c}(\mathbf{x})$. Generalizing to the case of a non-centered process Y with known trend function $\mu(.)$, we get the SK equations:

$$m_{\mathrm{SK}}(\mathbf{x}) = \mu(\mathbf{x}) + \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \tag{2}$$

$$s^2_{\mathrm{SK}}(\mathbf{x}) = C(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}), \tag{3}$$

where $\boldsymbol{\mu} = \mu(\mathbf{X})$ is the vector of trend values at the design of experiments. Classical properties include the fact that m_{SK} interpolates the data $(\mathbf{X}, \mathbf{y})$, and that s^2_{SK} is non-negative and zero at the design of experiments. Furthermore, the SK variance is independent of $\mathbf{y}$ (*homoscedasticity in the observations*). Note that these properties hold whatever the chosen kernel C, without any condition of compatibility with the data. In addition, let us remark that in the typical case of a stationary kernel $C(\mathbf{x}, \mathbf{x}') = \sigma^2 R(\mathbf{x} - \mathbf{x}'; \boldsymbol{\psi})$, the Simple Kriging equations simplify to $m_{\mathrm{SK}}(\mathbf{x}) = \mu(\mathbf{x}) + \mathbf{r}(\mathbf{x})'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and $s^2_{\mathrm{SK}}(\mathbf{x}) = \sigma^2\left(1 - \mathbf{r}(\mathbf{x})'\mathbf{R}^{-1}\mathbf{r}(\mathbf{x})\right)$, where $\mathbf{r}(\mathbf{x})$ and

$\mathbf{R}$ respectively stand for the analogues of $\mathbf{c}(\mathbf{x})$ and $\mathbf{C}$ in terms of correlation. In particular, m_{SK} is hence not depending on σ^2, while s^2_{SK} is proportional to it.

One major fact concerning the SK equations is that they coincide with classical conditioning results in the case where the process Z is assumed Gaussian. Indeed, the orthogonality of $Y(\mathbf{x}) - \boldsymbol{\lambda}^*(\mathbf{x})'Y(\mathbf{X})$ and $Y(\mathbf{X})$ ensures independence in the latter case, so that $m_{\mathrm{SK}}(\mathbf{x}) = \mathbb{E}[Y_{\mathbf{x}}|Y(\mathbf{X}) = \mathbf{y}]$. Similarly, s^2_{SK} then coincide with the conditional variance $\mathrm{Var}[Y_{\mathbf{x}}|Y(\mathbf{X}) = \mathbf{y}]$, so that the conditional law of $Y_{\mathbf{x}}$ can finally be written in terms of SK quantities:

$$Y_{\mathbf{x}}|Y(\mathbf{X}) = \mathbf{y} \sim \mathcal{N}(m_{\mathrm{SK}}(\mathbf{x}), s^2_{\mathrm{SK}}(\mathbf{x})) \tag{4}$$

More generally, the law of the whole random process Y conditional on $Y(\mathbf{X}) = \mathbf{y}$ is gaussian with trend m_{SK}, and with a conditional covariance struture which can be analytically derived in the same fashion as s^2_{SK} (see Ginsbourger *et al.* (2010) for details). The latter is the key to conditional simulations of Y knowing the observations, as we will illustrate section 4. We now turn to the case where the trend μ is known up to a set of linear trend coefficients.

When some linear trend coefficients are unknown: Universal Kriging

Let us focus on the case where the trend is of the form $\mu(\mathbf{x}) = \sum_{j=1}^{l} \beta_j f_j(\mathbf{x})$ ($l \in \mathbb{N} - \{0\}$), where the f_j's are fixed basis functions, and the β_j's are unknown real coefficients. Universal Kriging (UK) consists in deriving best linear predictions of Y based on the observations $Y(\mathbf{X})$ while estimating the vector $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_l)$ on the fly. The UK equations are given by:

$$m_{\mathrm{UK}}(\mathbf{x}) = \mathbf{f}(\mathbf{x})'\widehat{\boldsymbol{\beta}} + \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}) \tag{5}$$

$$s^2_{\mathrm{UK}}(\mathbf{x}) = s^2_{\mathrm{SK}}(\mathbf{x}) + (\mathbf{f}(\mathbf{x})' - \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}\mathbf{F})'(\mathbf{F}'\mathbf{C}^{-1}\mathbf{F})^{-1}(\mathbf{f}(\mathbf{x})' - \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}\mathbf{F}) \tag{6}$$

where $\mathbf{f}(\mathbf{x})$ is the vector of trend functions values at $\mathbf{x}$, $\mathbf{F} = (\mathbf{f}(\mathbf{x}^{(1)})', \ldots, \mathbf{f}(\mathbf{x}^{(n)})')'$ is the $n \times l$ so-called *experimental matrix*, and the best linear estimator of $\boldsymbol{\beta}$ under correlated residual is given by the usual formula $\widehat{\boldsymbol{\beta}} := (\mathbf{F}'\mathbf{C}^{-1}\mathbf{F}')^{-1}\mathbf{F}'\mathbf{C}^{-1}\mathbf{y}$. Basic properties of UK include similar interpolating behaviour as SK, with a variance vanishing at the design of experiments. Furthermore, $m_{\mathrm{UK}}(\mathbf{x})$ tends to the best linear fit $\mathbf{f}(\mathbf{x})'\widehat{\boldsymbol{\beta}}$ whenever the covariances $\mathbf{c}(\mathbf{x})$ vanish, which may typically happen when $\mathbf{x}$ is far away from the design $\mathbf{X}$ for some norm $||.||$, in the case of a stationary covariance kernel decreasing with $||\mathbf{u} - \mathbf{v}||$. Note also the inflation of the Kriging variance term, which reflects the additional uncertainty due to the estimation of $\boldsymbol{\beta}$.

As in the case of SK, it is possible to interprete UK in terms of random process conditioning, at the price of some specific assumptions. Indeed, working in a Bayesian framework with an improper uniform prior over $\mathbb{R}^p$ for $\boldsymbol{\beta}$ (Cf. Helbert, Dupuy, and Carraro (2009)) leads to a gaussian posterior distribution for the process Y conditional on the observations. Again, $m_{\mathrm{UK}}(\mathbf{x})$ and $s^2_{\mathrm{UK}}(\mathbf{x})$ appear respectively as conditional mean and variance, and the analytically tractable conditional covariance kernel enables the use of conditional simulations at any set of new design points. This model is a first step towards Bayesian Kriging, where more general prior distributions can be chosen, not only for $\boldsymbol{\beta}$ but also for all kernel parameters such as σ^2 or $\boldsymbol{\psi}$. We won't develop this approach any further here since a generic version of Bayesian Kriging is not proposed in the present version of **DiceKriging**, but send the interested reader to the seminal article Omre (1987) and the works of O'Hagan.

2.2. Filtering heterogeneously noisy observations with Kriging

In many practical situations, it is not possible to get exact evaluations of the deterministic

function y at the design of experiments, but rather punctual noisy measurements. This is the case for instance for PDE solvers relying on Monte-Carlo methods (e.g. in nuclear safety), or in partially converged simulations based on finite elements (e.g. in fluid dynamics). In such cases, for a given $\mathbf{x} \in D$, the user doesn't have access to $y(\mathbf{x})$, but to an approximate response $y(\mathbf{x})+\epsilon$. When it is reasonable to assume that ϵ is one realization of a "noise" random variable ε, it is still possible to derive Kriging approximations. Here we assume that the probability distribution of ε may depend on $\mathbf{x}$ and other variables, and that its realizations may differ for different measurements of y at the same $\mathbf{x}$. So instead of referring to the measurements of y in terms of $\mathbf{x}$'s, we will denote by $\tilde{y_i} = y(\mathbf{x}^i) + \epsilon_i$ the sequence of noisy measurements, where the $\mathbf{x}^i$'s are not necessarily all distinct, and by τ_i^2 the corresponding noise variances. Following the convenient particular case of Monte-Carlo simulations, we finally make the assumption that $\varepsilon_i \sim \mathcal{N}(0, \tau_i^2)$ $(1 \leq i \leq n)$ independently. Note that although not explicitly addressed in this paper, the case of multigaussian vector of ε_i's with prescribed covariance matrix is a straightforward generalization of the model considered here. We now recall the Kriging equations for heterogeneously noisy observations, in the Gaussian Process framework.

If we suppose that y is a realisation of a GP following the Simple Kriging assumptions above, the $\tilde{y_i}$'s can now be seen as realizations of the random variables $\tilde{Y_i} := Y(\mathbf{x}^i) + \varepsilon_i$, so that Kriging amounts to conditioning Y on the heterogeneously noisy observations $\tilde{Y_i}$ $(1 \leq i \leq n)$. Indeed, provided that the process Y and the gaussian measurement errors ε_i are stochastically independent, the process Y is still gaussian conditionally on the noisy observations $\tilde{Y_i}$ $(1 \leq i \leq n)$, and its conditional mean and variance functions are given by the following slightly modified Kriging equations:

$$m_{\mathrm{SK}}(\mathbf{x}) = \mu(\mathbf{x}) + \mathbf{c}(\mathbf{x})'(\mathbf{C} + \mathbf{\Delta})^{-1}(\widetilde{\mathbf{y}} - \boldsymbol{\mu}) \quad (7)$$

$$s^2_{\mathrm{SK}}(\mathbf{x}) = C(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})'(\mathbf{C} + \mathbf{\Delta})^{-1}\mathbf{c}(\mathbf{x}), \quad (8)$$

where $\widetilde{\mathbf{y}} = (\tilde{y_1}, \ldots, \tilde{y_n})'$, and Δ is the diagonal matrix of diagonal terms $\tau_1^2 \ldots \tau_n^2$. The only difference compared to noiseless SK equations is the replacement of C by $C + \Delta$ at every occurence. Specific properties of this variant of the SK metamodel include the fact that $m_{\mathrm{SK}}(.)$ is not interpolating the noisy observations (i.e. where no observation has been done with $\tau = 0$), that $s^2_{\mathrm{SK}}(.)$ doesn't vanish at that points, and is globally inflated compared to the noiseless case. Note that although $s^2_{\mathrm{SK}}(.)$ now depends on both the design $\mathbf{X}$ and the noise variances $\boldsymbol{\tau}^2 := \{\tau_1^2, \ldots, \tau_n^2\}$, it still doesn't depend on the observations, similarly as in the noiseless case. Note finally that the same filtering effect applies in the case of Universal Kriging, where the equations are similarly obtained by replacing $\mathbf{C}^{-1}$ by $(\mathbf{C} + \mathbf{\Delta})^{-1}$.

2.3. Covariance kernels and related parameter estimation

The choice of C has crucial consequences on the Kriging metamodel obtained, all the more so when the trend is known or assumed constant. In order to be admissible, C has to be chosen in the set of positive definite kernels. Checking that positive-definiteness holds for a given C is however not an easy task, and non-parametric estimation of it seem totally unrealistic. So what one typically does in Kriging is to select beforehand a parametric family of kernels known to be positive definite, and to estimate the corresponding parameters based on available data, for instance by maximizing a likelihood or minimizing the average cross-validation error. A usual restrictionis to consider kernels depending only on the increment $\mathbf{u}-\mathbf{v}$, called *stationary kernels*. Admissible stationary kernels coincide with the *functions of positive type*, which are

characterized by Bochner's theorem Rasmussen and Williams (2006) as Fourier transforms of positive measures. Some of the most popular 1-dimensional stationary kernels include the gaussian kernel, Fourier transform of the gaussian density, as well as the Matérn kernel, Fourier transform of the Student density. One convenient way of getting admissible covariance kernels in higher dimensions is to take tensor products of 1-d admissible kernels. Such kernels, called *separable*, are the most widely used in computer experiments literature. The covariance kernels available in the current version of **DiceKriging** are build upon this model, up to a multiplicative constant $\sigma^2 > 0$, with the 1-d functions summed up in Tab. 2.3 hereunder.

$$c(\mathbf{h}) := C(\mathbf{u}, \mathbf{v}) = \sigma^2 \prod_{j=1}^{d} g(h_j; \boldsymbol{\theta}_j), \tag{9}$$

where $\mathbf{h} = (h_1, \ldots, h_d) := \mathbf{u} - \mathbf{v}$. Although this could make sense in some contexts, the package does not allow mixing different covariance functions by dimensions. The parameters $\boldsymbol{\theta}_j's$ are chosen to be physically interpretable in the same unit as the corresponding variables. They are called *characteristic length-scales* by Rasmussen and Williams (2006). The analytic formula are taken from this book, and are given below (where $\theta > 0$ and $0 < p \leq 2$):

Gaussian	$g(h) = \exp\left(-\frac{h^2}{2\theta^2}\right)$
Matérn $\nu = 5/2$	$g(h) = \left(1 + \frac{\sqrt{5}\lvert h\rvert}{\theta} + \frac{5h^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}\lvert h\rvert}{\theta}\right)$
Matérn $\nu = 3/2$	$g(h) = \left(1 + \frac{\sqrt{3}\lvert h\rvert}{\theta}\right) \exp\left(-\frac{\sqrt{3}\lvert h\rvert}{\theta}\right)$
Exponential	$g(h) = \exp\left(-\frac{\lvert h\rvert}{\theta}\right)$
Power-Exponential	$g(h) = \exp\left(-\left(\frac{\lvert h\rvert}{\theta}\right)^p\right)$

The above covariances will result in different level of smoothness for the associated random process paths. With gaussian covariance, the sample paths of the associated centered GP have derivatives at all orders and are thus very smooth (they are even analytical). With Matérn covariance with parameter ν, the process is (mean square) differentiable at order k if and only if $\nu > k$. Thus with $\nu = 5/2$, the process is twice differentiable; with $\nu = 3/2$ only once. With $\nu = 1/2$, equivalent to the exponential covariance, the process is only continuous. This is also the case for the Power-Exponential covariance when the power parameter p is strictly less than 2. The general Matérn covariance depends on the modified Bessel function, and has not been implemented yet. However, when $\nu = k + 1/2$ where k is a non-negative integer, the covariance expression is given by an analytical expression. Two cases provided correspond to commonly needed levels of smoothness ($\nu = 3/2$ and $\nu = 5/2$).

Despite the offered flexibility in terms of differentiability level, all the covariance kernels above correspond to GPs with continous paths. Now, in applications, the assumption of continuity is sometimes untenable for simulator outputs, although deterministic, due in particular to numerical unstabilities. Hence, even if several evaluations at the same point deliver the same response, a slight change in the input vector may sometimes result in a jump in the response. Such discontinuities are classically handled in geostatistics using a so-called *nugget effect*,

consisting in adding a constant term τ^2 (named *jitter* in machine learning) to $c(\mathbf{0})$:

$$c(\mathbf{h}) := \sigma^2 \prod_{j=1}^{d} g(h_j; \boldsymbol{\theta}_j) + \tau^2 \delta_{\mathbf{0}}(\mathbf{h}) \tag{10}$$

The consequences of such a modification of the covariance kernel on Kriging are fairly similar to the case of noisy observations. Up to a rescaling of the Kriging variance due to the fact that the process variance changes from σ^2 to $\sigma^2 + \tau^2$, predicting with nugget or noisy observations coincide when considering points outside of the design of experiments: a diagonal with constant term is added to the covariance matrix of observations, smoothing out the noiseless Kriging interpolator and inflating the variance. A major difference, however, is that Kriging with nugget effect conserves the interpolation property: since τ^2 appears this time in the covariance vector too, the Kriging mean predictor systematically delivers the original observation in virtue of the covariance vector being a column of the covariance matrix, as in the deterministic case without nugget and contrarily to the noisy case.

In **DiceKriging**, the parameters can be either given by the user or estimated. The first situation is useful for academic research or for Bayesian computations. The second one for non-Bayesian approaches, and for kriging-based optimization. At present time, two estimation methods are proposed: Maximum Likelihood Estimation (MLE) or Penalized MLE (PMLE). The latter is based on the SCAD penalty defined in Li and Sudjianto (2005) but is still in beta version. In appendix A, we give the expressions of the likelihoods, concentrated likelihoods and analytical derivatives involved. We also report to section 3.3 for optimization details.

2.4. Kriging-based Optimization: Expected Improvement and EGO

Optimization (say minimization) when the objective function is evaluated through costly simulations creates a need for specific strategies. In most cases indeed, the non-availability of derivatives prevents one from using gradient-based techniques. Similarly, the use of metaheuristics (e.g. genetic algorithms) is compromised by severely limited evaluation budgets.

Kriging metamodels has been sucessfully used for the global optimization of costly deterministic functions since the nineties Jones *et al.* (1998). A detailed review of global optimization methods relying on metamodels can be found in Jones (2001). The latter illustrates why directly minimizing a deterministic metamodel (like a spline, a polynomial, or the kriging mean) is not efficient. Kriging-based sequential optimization strategies address the issue of converging to non optimal points by taking the kriging variance term into account, hence incenting the algorithms to be more explorative. Such algorithms produce one point at each iteration that maximizes a figure of merit based upon the law of $Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{y}$. Several infill sampling criteria are available, that balance kriging mean prediction and uncertainty. The Expected Improvement criterion has become one of the most popular such criteria, probably thanks to both its well-suited properties and its analytical tractability.

The Expected Improvement criterion

The basic idea underlying EI is that sampling at a new point $\mathbf{x}$ will bring an improvement of $\min(y(\mathbf{X})) - y(\mathbf{x})$ if $y(\mathbf{x})$ is below the current minimum, and 0 otherwise. Of course, this quantity cannot be known in advance since $y(\mathbf{x})$ is unkown. However, the GP model and the

available information $Y(\mathbf{X}) = \mathbf{y}$ make it possible to define and derive:

$$\begin{aligned} \mathrm{EI}(\mathbf{x}) : &= \mathbb{E}\left[(\min(Y(\mathbf{X})) - Y(\mathbf{x}))^+ \,|Y(\mathbf{X}) = \mathbf{y}\right] \\ &= \mathbb{E}\left[(\min(\mathbf{y}) - Y(\mathbf{x}))^+ \,|Y(\mathbf{X}) = \mathbf{y}\right], \end{aligned} \tag{11}$$

for which an integration by parts yields the analytical expression (Cf. Jones *et al.* (1998)):

$$\mathrm{EI}(\mathbf{x}) := (\min(\mathbf{y}) - m(\mathbf{x}))\,\Phi\left(\frac{\min(\mathbf{y}) - m(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{\min(\mathbf{y}) - m(\mathbf{x})}{s(\mathbf{x})}\right), \tag{12}$$

where Φ and ϕ are respectively the cdf and the pdf of the standard gaussian law. The latter analytical expression is very convenient since it allows fast evaluations of EI, and even analytical calculation of its gradient and higher order derivatives. This used in particular in **DiceOptim** for speeding up EI maximization. This criterion has important properties for sequential exploration: it is null at the already visited sites, and positive everywhere else with a magnitude that is increasing with $s(.)$ and decreasing with $m(.)$.

The "Efficient Global Optimization" Algorithm

EGO (Cf. Jones *et al.* (1998)) relies on the EI criterion. Starting with an initial Design $\mathbf{X}$ (typically a Latin Hypercube), EGO sequentially visits a current global maximizer of EI and updates the Kriging metamodel at each iteration, including hyperparameters re-estimation:

1. Evaluate y at $\mathbf{X}$, set $\mathbf{y} = y(\mathbf{X})$, and estimate covariance parameters of Y by ML
2. While stopping criterion not met
   (a) Compute $\mathbf{x}^{\mathrm{new}} = argmax_{\mathbf{x}\in D}\mathrm{EI}(\mathbf{x})$ and set $\mathbf{X} = \mathbf{X} \cup \{\mathbf{x}^{\mathrm{new}}\}$
   (b) Evaluate $y(\mathbf{x}^{\mathrm{new}})$ and set $\mathbf{y} = \mathbf{y} \cup \{y(\mathbf{x}^{\mathrm{new}})\}$
   (c) Re-estimate covariance parameters by MLE and update Kriging metamodel

EGO and related EI algorithm have become commonplace in computer experiments, and are nowadays considered as reference global optimization methods in dimension $d \leq 10$ in cases where the number of objective function evaluations is drastically limited (See e.g. Jones (2001) or Brochu *et al.* (2009)). One major drawback of EGO is that is does not allow parallel evaluations of y, which is desirable for costly simulators (e.g. a crash-test simulation run typically lasts 24 hours). This was already pointed out in Schonlau (1997), where the multi-points EI was defined but not further developed. This work was continued in Ginsbourger *et al.* (2010) by expliciting the latter multi-points EI (q-EI), and by proposing two classes of heuristics strategies meant to approximatly optimize the q-EI, and hence simultaneously deliver an arbitrary number of points without intermediate evaluations of y.

2.5. Adaptations of EI and EGO for Synchronous Parallel Optimization

Considering $q \geq 2$ candidate design points $\mathbf{X}^{\mathrm{new}} := \{\mathbf{x}^{n+1)}, \ldots, \mathbf{x}^{(n+q)}\}$, the q-points EI criterion is straightforwardly defined as conditional expectation of the joint improvement brought by the new points:

$$\mathrm{EI}(\mathbf{x}^{(\mathrm{n+1})}, ..., \mathbf{x}^{(\mathrm{n+q})}) = \mathbb{E}\left[\left(\min(\mathbf{y}) - \min\left(\mathrm{Y}(\mathbf{x}^{(\mathrm{n+1})}), ..., \mathrm{Y}(\mathbf{x}^{(\mathrm{n+q})})\right)\right)^+ |\mathrm{Y}(\mathbf{X}) = \mathbf{y}\right] \tag{13}$$

Unlike in the 1-point situation, q-EI is not known in closed form (See Ginsbourger *et al.* (2010) for a formula in the case $q = 2$). However, it is possible to estimate it by a standard Monte Carlo technique relying on gaussian vector simulation:

function Q-$\mathbb{EI}(\mathbf{X}, \mathbf{y}, \mathbf{X}^{\text{new}})$
  $L = \text{chol}(Var[Y(\mathbf{X}^{new})|Y(\mathbf{X}) = \mathbf{y}])$ ▷ Decomposition of cond. cov. mat.
  for $i \leftarrow 1, n_{sim}$ **do**
    $N \sim \mathcal{N}(0, I_q)$ ▷ Drawing a standard gaussian vector N at random
    $M_i = m(\mathbf{X}^{new}) + LN$ ▷ Simulating Y at $\mathbf{X}^{new}$ conditional on $Y(\mathbf{X}) = \mathbf{y}$
    $qI_{sim}(i)$=$[\min(\mathbf{y}) - \min(M_i)]^+$ ▷ Simulating the improvement at $\mathbf{X}^{\text{new}}$
  end for
  $qEI_{sim} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} qI_{sim}(i)$ ▷ Estimation of the q-points Expected Improvement
end function

q-EI can potentially be used to deliver an additional design of experiments in one step through the resolution of the optimization problem

$$(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+2)}, ..., \mathbf{x}^{(n+q)}) = argmax_{\mathbf{X}^{\text{new}} \in D^q}[\text{EI}(\mathbf{X}^{\text{new}})] \tag{14}$$

However, the optimization problem defined by Eq. 14 is of dimension $d \times q$, and with a noisy and derivative-free objective function in the case where the criterion is estimated by Monte-Carlo. Pseudo-sequential greedy strategies have been proposed that approach the result of problem 14 while avoiding its numerical cost, hence circumventing the curse of dimensionality. In particular, the *Constant Liar* (CL) is a sequential strategy in which the metamodel is updated (still without hyperparameter re-estimation) at each iteration with a value $L \in \mathbb{R}$ exogenously fixed by the user, here called a "lie":

Algorithm 1 The Constant Liar algorithm:

function CL($\mathbf{X}$, $\mathbf{y}$, L, q)
  for $i \leftarrow 1, q$ **do**
    $\mathbf{x}^{n+i} = argmax_{\mathbf{x} \in D} \text{EI}(\mathbf{x})$
    $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^{n+i}\}$
    $\mathbf{y} = \mathbf{y} \bigcup \{L\}$
  end for
end function

The effect of L on the performance of the resulting optimizer is investigated in the next section. L should logically be determined on the basis of the values taken by y at $\mathbf{X}$. Three values, $min\{\mathbf{y}\}$, $mean\{\mathbf{y}\}$, and $max\{\mathbf{y}\}$ were considered in Ginsbourger *et al.* (2010). In the present version of **DiceOptim**, the `CL` function has a tunable L, whereas the parallel version of EGO relying on CL (called `CL.nsteps`) has a lie L fixed to $min\{\mathbf{y}\}$. More details are given in section 5, dedicated to numerical examples produced with the maine functions of **DiceOptim**.

3. Guidelines for users

3.1. The important functions

DiceKriging performs estimation, simulation, prediction and validation for various Kriging

models. The first step is to define a Kriging model, which is the goal of the `km` function. It is suited either for Simple Kriging (SK) and Universal Kriging (UK), noise-free and noisy observations, and allows some flexibility in estimating all or only some parameters. Its functioning is detailed in table 1. Simulation, prediction, and validation are implemented as `simulate`, `predict` and `plot` methods, that apply directly to `km` objects. For prediction, the kind of Kriging must be indicated in the argument `type` (`"SK"` or `"UK"`). The `plot` method corresponds to leave-one-out validation. A k-fold validation can be found in **DiceEval**.

DiceOptim performs sequential and parallel Kriging-based optimization, based on the 1-point and multipoints expected improvement criteria. The main functions are described in table 2. Note that in the version presented here, **DiceOptim** is limited to noise-free observations.

Kriging model description	**Arguments to be specified in `km`**
Unknown trend and cov. kernel parameters (UK)	(Default) Nothing to specify
Known trend and cov. kernel parameters (SK)	`coef.trend`, `coef.cov`, `coef.var`
Known trend, unknown cov. kernel parameters	`coef.trend`
Optional nugget effect - Known - Unknown	 `nugget` *(homogeneous to a variance)* `nugget.estim=TRUE`
Case of noisy observations *(incompatible with a nugget effect in this package)*	`noise.var` *(homogen. to a variance)*

Table 1: Possibilities of the function `km` for the definition of Kriging models. `km` estimates the unknown parameters and creates the resulting `km` object. Default is Universal Kriging (noise-free observations, no nugget effect). The other Kriging models are obtained by specifying arguments in `km` as indicated below.

R function	**Description**
`EI`	One-point noise-free EI criterion
`qEI`	q-points noise-free EI criterion (estimated by Monte Carlo)
`EGO.nsteps`	Sequential EI Algorithm – model updates including re-estimation of covariance parameters – with a fixed number of iterations (nsteps)
`max_EI`	One-point noise-free EI maximization. No call to the objective function
`max_qEI.CL`	(sub-)maximization of the q-points EI, based on the Constant Liar heuristic. No call to the objective function
`CL.nsteps`	Parallel EI Algorithm – model updates including re-estimation of covariance parameters – with a fixed number of iterations (nsteps)

Table 2: Important functions in **DiceOptim**

3.2. Trend definition and data.frames

A convenient way to specify linear trends in R is to use the class `formula` that provides compact symbolic descriptions. Among advantages of formulas, any functional term can be given, and updating models is very easy. For instance, the linear model arguments of the functions `lm` or `glm` (package **stats**) are given as objects of class `formula`. The same is possible in **DiceKriging**, through the `km` function, with some specificities that we describe now.

First, the design $\mathbf{X}$ must be provided as a `data.frame` in the argument `design` of `km`, which

implies that every columns of $\mathbf{X}$ are named. Then the trend can be specified in the argument `formula`, using these names. The `formula` mechanism mainly works as in the `lm` function from the **stats** package, as shown in table 3; in particular, remark that the inhibition function `I` is sometimes required, especially for polynomial terms. Note however that the left hand term is not needed (and will not be used if provided): this is because `formula` is only defining a trend (and not a transformation of the response $\mathbf{y}$).

Once the trend is specified, one can estimate or build a `km` object (see last section). For prediction (or simulation), the new location(s) $\mathbf{X}^{\text{new}}$ have to be specified in the argument `newdata` of `predict` (or `simulate`) method. In practice, the user may not want to loose time converting matrices to data.frames. Thus, $\mathbf{X}^{\text{new}}$ can be provided simply as a matrix (or even a vector in case of a single new data). However, with this practical syntax, no variable name is stored for $\mathbf{X}^{\text{new}}$, and it is *always* assumed that the columns of $\mathbf{X}^{\text{new}}$ are provided in the same order as for $\mathbf{X}$. For evident reasons, the user must not depart from this rule.

Trend	**`formula` in `km`**
Constant (default model)	`~ 1`
Full 1st order polynomial	`~ .`
idem + 2nd order interactions	`~ .^2`
$\beta_0 + \beta_5 x_5^3 + \beta_{2,6} x_2 x_6 + \beta_{12} x_{12}$	`~ I(x5^3) + I(x2*x6) + x12`
$\beta_0 + \beta_4 \cos(x_4) + \beta_7 \sin(x_7)$	`~ cos(x4) + sin(x7)`
idem, but without intercept	`~ -1 + cos(x4) + sin(x7)`
1st order polynomial without x_3	`~ . - x3`
1st order polynomial plus $\beta_9 \exp(x_8)$	`~ . + exp(x8)`
Full 2nd order polynomial $d = 3$	`~ .^2 + I(x1^2) + I(x2^2) + I(x3^2)`

Table 3: Examples of trend specification in `km`.

3.3. Comments about the optimization algorithms

Optimization algorithms are used on two different occasions: likelihood maximization and EI maximization. These problems are quite hard, due to the cost of objective functions, numerical instabilities, multimodality and dimensionality issues (see e.g. Santner *et al.* (2003)). In **DiceKriging** and **DiceOptim**, we have addressed them by following four leading ideas. First, rely on trustworthy state-of-the-art algorithms; Second, choose at least one algorithm capable to overcome multimodality; Third, to improve speed and accuracy, supply the analytical gradients; Fourth, enable tunability of key parameters by the user. For these reasons, we have selected a quasi-Newton BFGS algorithm (function `optim`) and the hybrid algorithm `genoud` (from package **rgenoud**), taking advantage of both the global search provided by a genetic algorithm and the local search based on gradients. Let us now turn to some specificities.

Likelihood maximization. The implementation is based on the efficient algorithm proposed by Park and Baek (2001), and extended to the Kriging models addressed here (see the implementation details in the appendix A). On the one hand, the results obtained with BFGS are the quickest, but may be variable. To reduce this variability, BFGS is run from the best point among an initial population drawn at random. The size of this population can be tuned by changing `pop.size` in the argument `control`. On the other hand, `rgenoud`

usually gives more stable results, at the cost of a higher computational time. Again, the most influential parameters can be tuned by the user. Some results concerning the performance of the optimizers are given in the examples section.

EI maximization. The branch-and-bound algorithm proposed by Jones *et al.* (1998) was not chosen here, since the boundaries depend on the covariance function and may be hard to find in general. Due to the multimodal form of EI (which is equal to 0 at every visited point), the `genoud` algorithm is used. The analytical gradient of EI can is derived in appendix B.

3.4. Known problems and their solutions

Despite the care taken in implementation, some problems can be encountered. Some of them are purely numerical, and can be solved easily, but others are intrinsic to Kriging and may be due to an inappropriate choice of design points or covariance kernels.

Non invertibility of covariance matrices. In some frequent situations, interpolation is hard to achieve numerically. Roughly speaking, this happens when the design points are too close relatively to the spatial correlation length. This results in nearly non invertible covariance matrices. The problem is all the more severe that the covariance kernel is constraining. With our kernels, the worst case with this respect is the Gaussian kernel, which implies the existence of derivatives at any order. On the other hand, the Matérn kernel with $\nu = 5/2$ (for which the derivatives exist up to order 2) gives much better conditioned covariance matrices. Thus, the first recommendation is to avoid using the Gaussian kernel and prefer the Matérn kernel with $\nu = 5/2$; Such a choice is strongly advocated by Stein (1999). This kernel is the default choice in **DiceKriging**. Another possibility, that can be combined to the former one, is to add a nugget effect, or *jitter*. This method is sometimes referred to as *diagonal inflation*. However, the sample paths are then discontinuous at the design points. Therefore, the nugget value should not be too large in order to avoid abusive departure from continuity, but not too small to prevent from numerical problems. An illustration is given in section 4.6.

Identifiability issues caused by large design interdistances. A dual difficulty is encountered when the design points are *not* close enough relatively to the spatial correlation length. In such situations, estimation of Kriging models may give either misleading results or *flat* predictions, corresponding to range parameters θ estimated to zero. Analogous issues are faced in signal theory, where recovering a periodic signal is not possible if the sampling frequency is too small. A solution is penalizing. For instance, Li and Sudjianto (2005) have shown some promising results obtained by Penalized MLE with SCAD penalty (Fan (1997)). This method has been implemented in **DiceKriging**, but should be considered as a beta version at this stage, since the estimation results for the tuning parameter are not convincing. Another possibility is simply to add a constraint $\theta \geq \theta_{min}$ in MLE. One then face the analogous problem of choosing the lower threshold θ_{min}. An illustration is given in section 4.6.

3.5. Trustworthiness

Producing a trustworthy result is the main concern – or "prime directive", as named by

Chambers (2008) – of software providers. To experiment trustworthiness of our packages, we have implemented several tests, some of them being included in the examples of functions `simulate.km` and `km`. The first one is a consistency test between the prediction formulas of Simple Kriging and simulation results; in limit cases, prediction with Universal Kriging is also successfully compared to the linear regression confidence bounds computed with the `lm` [**stats**] function. The second one is a performance study of maximum likelihood estimators, achieved by monitoring Kriging parameters estimates based on simulated paths of a gaussian process which law is actually known. This study is higlighted in one of the examples of section 4.

4. Examples with DiceKriging

4.1. An introductory 1D example with known parameters.

First, let use `km` to build a kriging model with the following characteristics: 2nd order polynomial trend $-1 + 2x - 0.5x^2$, Matérn 5/2 covariance structure with $\sigma = 3$ and $\theta = 0.4$. Recall that the trend form is interpreted with an object of type `formula`, in the same way as for linear models (see `lm{stats}`). As this formula is based on the variables names, the argument `design` of `km` must be a `data.frame`. Thus in the following example, the vector `inputs` is converted to a `data.frame` with name `x`, which must also be used in the argument `formula`.

```
> inputs <- c(-1, -0.5, 0, 0.5, 1); output <- c(-9, -5, -1, 9, 11)
> theta <- 0.4; sigma <- 5; trend <- c(0,11,2);
> model <- km(formula=~x+I(x^2), design=data.frame(x=inputs), response=output,
+ covtype="matern5_2", coef.trend=trend, coef.cov=theta, coef.var=sigma^2)
```

Note that for polynomials, the operator `I` must be used in `formula`. Mathematically, the resulting quadratic trend is $11x + 2x^2$. For nice printing, and checking purpose, just type:

```
> model

Call:
km(formula = ~x + I(x^2), design = data.frame(x = inputs), response = output,
    covtype = "matern5_2", coef.trend = trend, coef.cov = theta,
    coef.var = sigma^2)

Trend  coeff.:

 (Intercept)     0.0000
           x    11.0000
      I(x^2)     2.0000

Covar. type  : matern5_2
Covar. coeff.:

    theta(x)     0.4000

Variance: 25
```

Then use the `predict` method to predict at newdata. As all parameters are assumed to be known, the argument `type` is turned to "SK", standing for Simple Kriging.

```
> t <- seq(from=-2, to=2, length=200)
> p <- predict(model, newdata=t, type="SK")
```

Finally plot the results : SK mean and 95% confidence intervals.

```
> plot(t, p$mean, type="l", xlim=c(-2,2), ylim=c(-30,30), xlab="x", ylab="y")
> lines(t, p$lower95, col="black", lty=2)
> lines(t, p$upper95, col="black", lty=2)
> points(inputs, output, col="red", pch=19)
> abline(h=0)
```

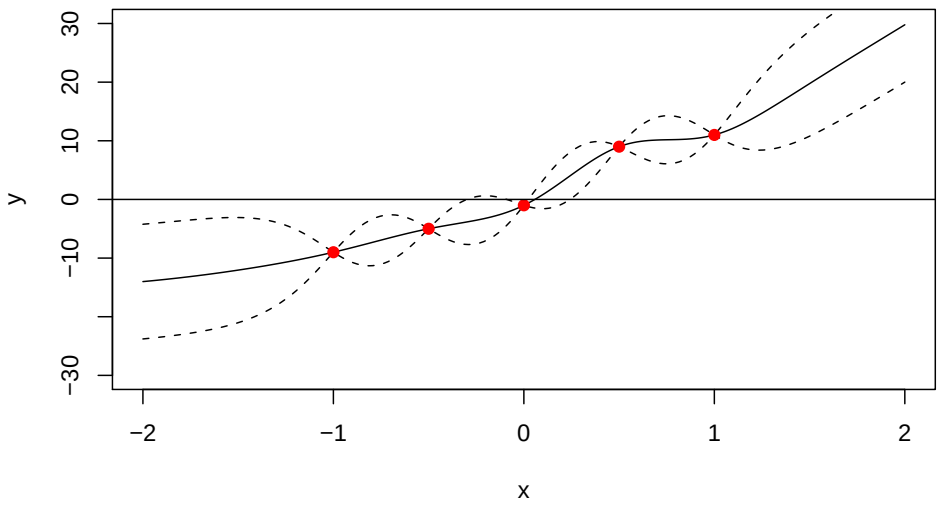

Figure 1: A first 1-dimensional example of Simple Kriging with second order polynomial trend and Matérn covariance. All parameters are known here.

Influence of the range parameters

The range parameters in **DiceKriging** are *length-scale*: a small value is analogue to a high frequency, and a large one to a low frequency. To illustrate this, we have run the code above with three values of θ: 0.05, 0.4 and 1. The result is represented in figure 2.

Influence of the trend

Let us visualize the influence of the trend, by comparing constant, affine and sine trends. The sine trend is chosen instead of the quadratic trend, to show that the trends proposed in **DiceKriging** are not limited to polynomials. The only modification in the R code is in the argument `formula`. For a constant trend, we have used:

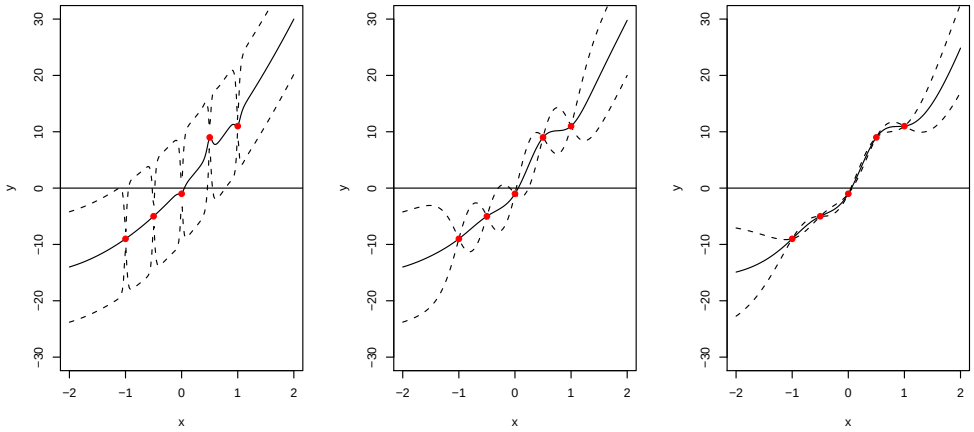

Figure 2: Simple Kriging with known parameters, and three different range values: small (left), intermediate (middle) and large (right). The range parameters are length-scale.

```
> formula <- ~1; trend <- 0
```

For the affine trend:

```
> formula <- ~x; trend <- c(0,10)
```

The first coefficient in `trend` is the intercept, and the second one the slope: thus, mathematically the affine trend is $10x$. Note that in this case, another possible choice for `formula` is:

```
> formula <- ~.
```

Finally, for the sine trend, we have used the function $1 + 15\sin(\frac{\pi}{4}x)$:

```
> formula <- ~sin(pi/4*x); trend <- c(1,15)
```

The corresponding results are shown in figure 3. The main difference is with extrapolation (here mainly outside $[-1, 1]$) where the Kriging mean reverts to the specified trend. For more details, see Ginsbourger, Dupuy, Badea, Carraro, and Roustant (2009).

4.2. Simulations of gaussian processes underlying Kriging models

Conditional and unconditional simulations are implemented as the so-called `simulate` method, applying to `km` objects. For instance, `simulate` can be applied to the object `model` defined in the previous section 4.1. By default, unconditional simulations are performed at design points; to simulate at new points, the new locations must be specified in the argument `newdata`. The argument `nsim` contains the number of simulations required.

```
> t <- seq(from=-2, to=2, length=200)
> y <- simulate(model, nsim=5, newdata=t)
```

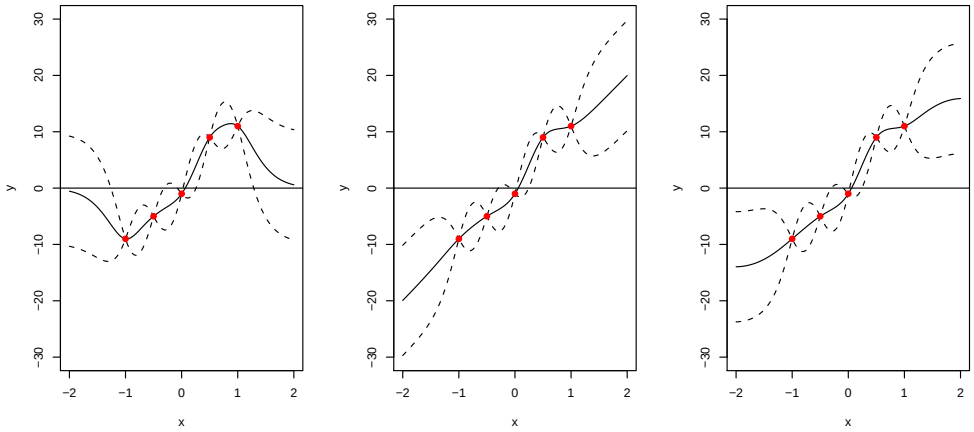

Figure 3: Simple Kriging with known parameters, and three different trends: constant, affine and sine.

Formally the value `y` is a matrix where each row contains one simulation. All simulations are represented in figure 4. The trend is the 2^d order polynomial previously considered.

```
> ytrend <- trend[1] + trend[2]*t + trend[3]*t^2
> plot(t, ytrend, type="l", col="black", ylab="y", lty="dashed",
+           ylim=c(min(ytrend)-2*sigma, max(ytrend) + 2*sigma))
> for (i in 1:5) lines(t, y[i,], col=i)
```

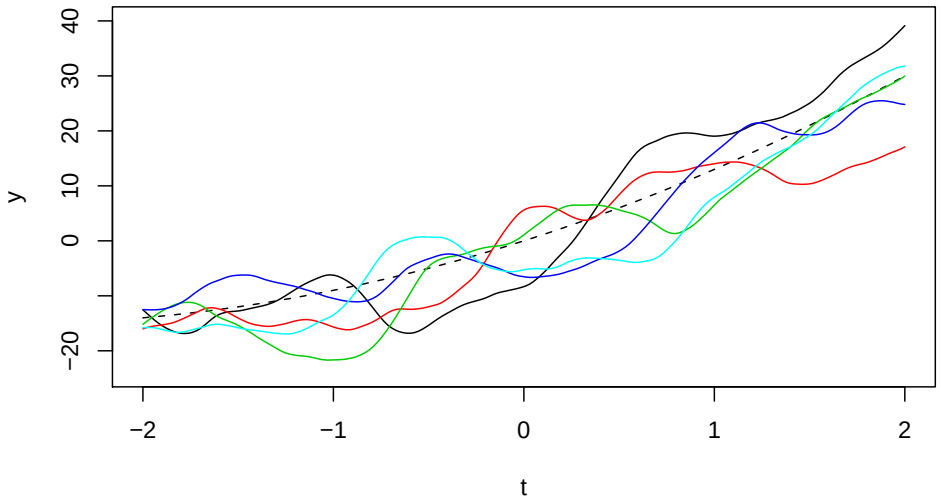

Figure 4: Unconditional simulations of a trended 1-d GP process with Matérn covariance

Influence of covariance functions

The smoothness of stationary GP sample functions depends on the properties (at **0**) of the covariance functions. To illustrate this, one simulation is performed for each covariance function `"exp"`, `"matern3_2"`, `"matern5_2"` or `"gauss"`. To save space, the code is not included in the vignette, but can be found in the documentation of `simulate.km`. The different simulations are represented in figure 5, as well as the corresponding covariance functions. The same kind of experiment can be done with the power-exponential covariance function (`"powexp"`), also implemented in **DiceKriging**.

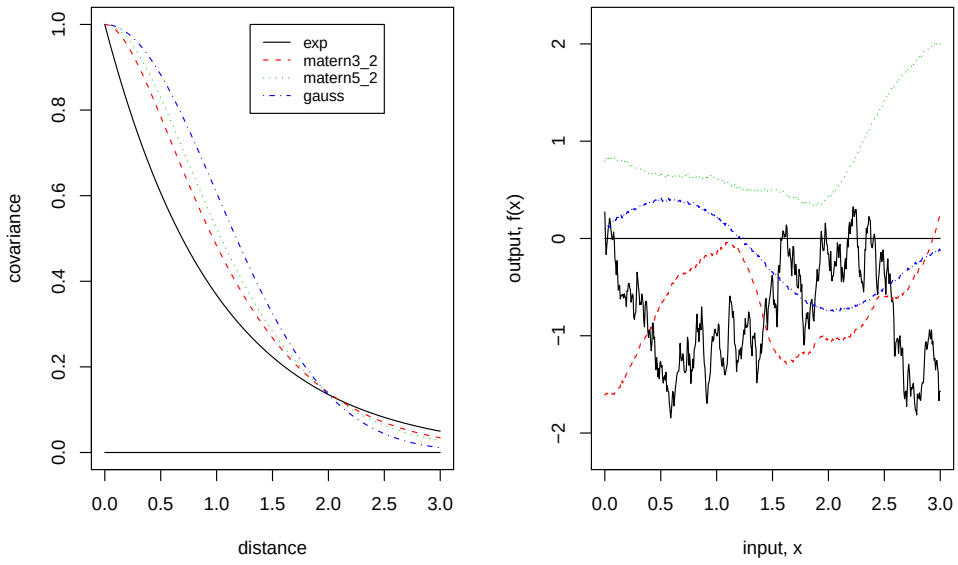

Figure 5: 1D simulations of GP processes

Conditional simulations

Still following the first example of section 4.1, conditional simulations can be performed, simply by turning the argument `cond` to TRUE:

```
> y <- simulate(model, nsim=5, newdata=t, cond=TRUE)
```

The conditional simulations are represented in figure 6. This procedure will be used in the next section to do two trustworthy tests. Firstly, the conditional simulations can be compared with the Simple Kriging mean and variance. Secondly, it will be used to evaluate the accuracy of the maximum likelihood estimators of the GP covariance parameters.

```
> ytrend <- trend[1] + trend[2]*t + trend[3]*t^2
> plot(t, ytrend, type="l", col="black", ylab="y", lty="dashed",
+           ylim=c(min(ytrend)-2*sigma, max(ytrend) + 2*sigma))
> for (i in 1:5) lines(t, y[i,], col=i)
> points(inputs, output, col="red", pch=19)
```

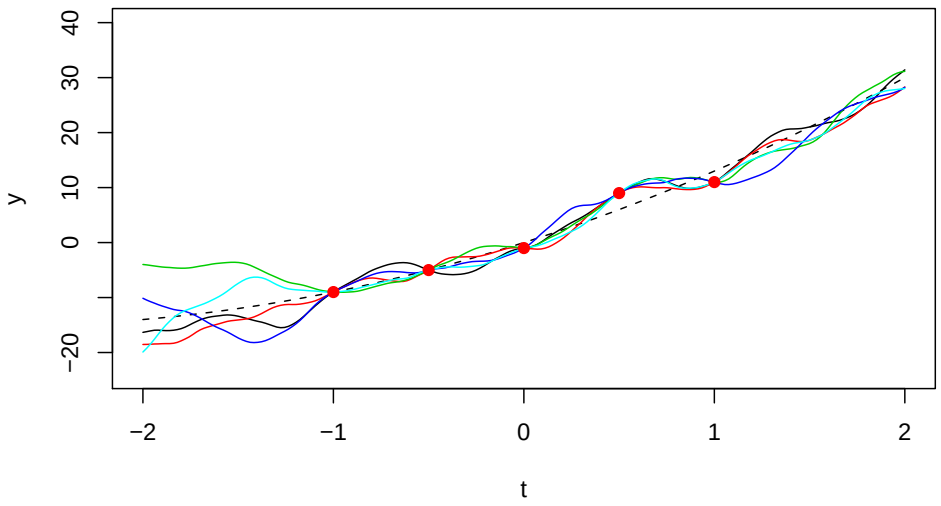

Figure 6: Conditional simulations of a kriging model.

4.3. Estimation and validation of Kriging models

A 2-dimensional case study

To illustrate the functionalities of **DiceKriging** on non-simulated data, we start with the famous 2-dimensional Branin-Hoo function. This function is a usual case study in global optimization with the particularity of having three global minimizers. Thus, it will be used in several **DiceOptim** examples. At now, let us estimate a kriging model. In this first example, we consider a very naive design, which is a 4×4 grid.

```
> X <- expand.grid(seq(0,1,length=4), seq(0,1,length=4))
> X <- data.frame(X); names(X)<-c("x1", "x2")
> y <- apply(X, 1, branin)
```

For the trend, we use a 1st order polynomial, but this choice does not influence a lot the results here. Finally, assuming that we assume *a priori* that the Branin-Hoo function is smooth, we use a Gaussian covariance function. Then, estimation is performed using `km`:

```
> m <- km(~., design=X, response=y, covtype="gauss")

optimisation start
------------------
* optimisation method : BFGS
* analytical gradient : used
* trend model : ~x1 + x2
* covariance model :
```

```
  - type :  gauss
  - nugget : NO
  - parameters lower bounds :  1e-10 1e-10
  - parameters upper bounds :  2 2
  - best initial point among  20  :  0.9750077 1.694222

N = 2, M = 5 machine precision = 2.22045e-16
At X0, 0 variables are exactly at the bounds
At iterate     0  f=       76.395  |proj g|=      0.97501
At iterate     1  f =       76.038  |proj g|=       0.90575
At iterate     2  f =        74.99  |proj g|=        1.2404
At iterate     3  f =       74.776  |proj g|=       0.68418
At iterate     4  f =       74.768  |proj g|=       0.12204
At iterate     5  f =       74.768  |proj g|=      0.005288
At iterate     6  f =       74.768  |proj g|=    4.3439e-05
At iterate     7  f =       74.768  |proj g|=    3.3386e-08

iterations 7
function evaluations 9
segments explored during Cauchy searches 9
BFGS updates skipped 0
active bounds at final generalized Cauchy point 1
norm of the final projected gradient 3.33856e-08
final function value 74.7675

F = 74.7675
final  value 74.767536
converged
```

Verbosity can be removed by means of the argument `control`,

```
> m <- km(~., design=X, response=y, covtype="gauss", control=list(trace=FALSE))
```

but there are several advantages to keep it, at least in this vignette, to see what is tunable in estimation. Indeed, many default values are proposed by `km`: the maximization of the likelihood is performed with the BFGS optimizer, using analytical gradient and an initial random search based on 20 initial points. Also, the domain over which the parameters are estimated, depending on the ranges of the design `X` in each direction. But in many cases, in can be interesting to change all these default values. For instance, the number of initial random search points can be reduced:

```
> m <- km(~., design=X, response=y, covtype="gauss",
+               control=list(pop.size=5, trace=FALSE))
```

The domain boundaries are contained in the arguments `lower` and `upper`. The genetic algorithm `rgenoud` usually gives better results, and is specified using the argument `optim.method`.

```
> m <- km(~., design=X, response=y, covtype="gauss", optim.method="gen",
+               control=list(trace=FALSE))
```

As for BFGS, the main optimization parameters such as the size of the initial population can be tuned in `control`. Coming back to the default values, we now print the estimation results:

```
> m

Call:
km(formula = ~., design = X, response = y, covtype = "gauss")

Trend  coeff.:
               Estimate
 (Intercept)  1249.2314
          x1  -672.2751
          x2  -362.5807

Covar. type  : gauss
Covar. coeff.:
               Estimate
   theta(x1)     0.8461
   theta(x2)     2.0000

Variance estimate: 855174.6
```

We can see a clear anisotropy with a longer range in the x_2 direction. The estimated value of θ_2 reached the boundary 2. Since it depends on the two ranges θ_1 and θ_2 only, the concentrated likelihood can be plotted. In the following, we compute `logLikFun` over a 30×30 grid:

```
> n.grid <- 30; x.grid <- seq(0.01,2,length=n.grid)
> X.grid <- expand.grid(x.grid, x.grid)
> logLik.grid <- apply(X.grid, 1, logLikFun, m)
```

The result can then be drawn, and the optimum added by extracting it from the `km` object `m`:

```
> contour(x.grid, x.grid, matrix(logLik.grid, n.grid, n.grid), 40,
+              xlab=expression(theta[1]), ylab=expression(theta[2]))
> opt <- m@covariance@range.val
> points(opt[1], opt[2], pch=19, col="red")
```

In figure 7, we see that the optimum is found in a correct way by BFGS. This is because the likelihood surface is pretty simple. Of course, this may *not* be always the case, especially in higher dimensions. Tuning the argument `optim.method` to `"gen"` may be useful.

Now, we can draw the kriging mean and visualize the prediction accuracy.

```
> n.grid <- 50
> x.grid <- seq(0,1,length=n.grid)
> X.grid <- expand.grid(x.grid, x.grid)
> y.grid <- apply(X.grid, 1, branin)
> pred.m <- predict(m, X.grid, "UK")
```

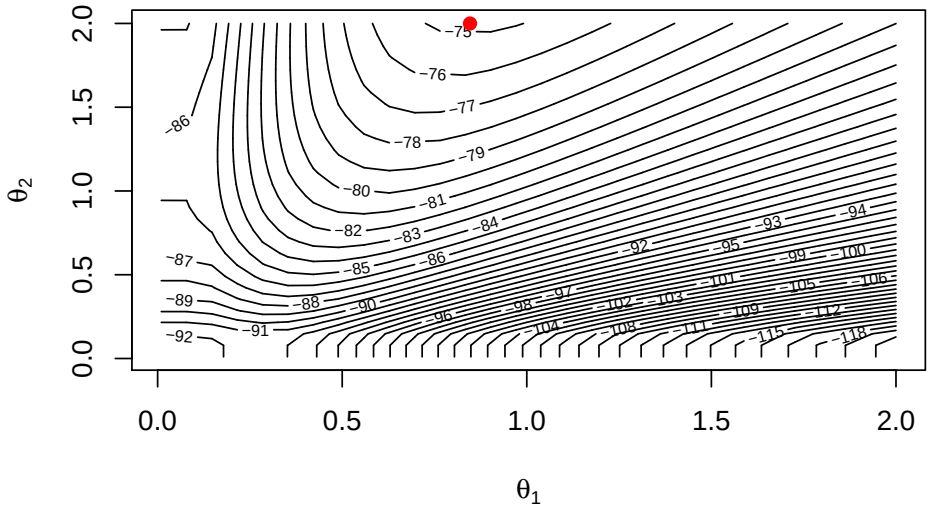

Figure 7: Profile log-likelihood

```
> par(mfrow=c(1,3))
> contour(x.grid, x.grid, matrix(y.grid, n.grid, n.grid), 50, main="Branin")
> points(X[,1], X[,2], pch=19, cex=1.5, col="red")
> contour(x.grid, x.grid, matrix(pred.m$mean, n.grid, n.grid), 50,
+              main="Kriging mean")
> points(X[,1], X[,2], pch=19, cex=1.5, col="red")
> contour(x.grid, x.grid, matrix(pred.m$sd^2, n.grid, n.grid), 15,
+              main="Kriging variance")
> points(X[,1], X[,2], pch=19, cex=1.5, col="red")
```

The kriging variance is also represented in figure 8. We observe that the prediction is satisfactory. The anisotropy appears clearly on the graph of the kriging variance.

Finally, a leave-one-out validation is implemented as a `plot` method (see results in figure 9):

```
> plot(m)
```

A more complete validation procedure is available in package **DiceEval** (not presented here), including k-fold cross validation.

A 6-dimensional approximation example

Let us now consider the 6-dimensional Hartman function. It is a standard test function in optimization, and is implemented in **DiceKriging**. Following Jones *et al.* (1998), the transformation $-log(-.)$ is first applied to the response. For estimation, a 80-point LH design is chosen at random (function `randomLHS` from package **lhs**). The leave-one-out diagnostic is represented in figure 10, and shows no strong departures from the model assumptions.

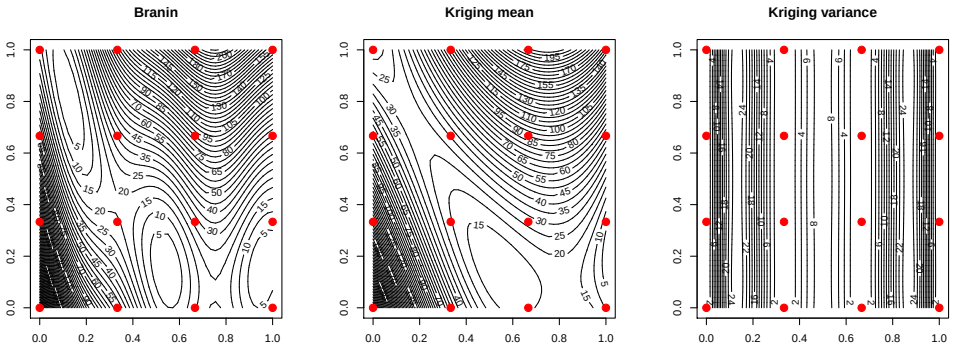

Figure 8: Contours of the Branin-Hoo function (left) and Ordinary Kriging metamodel of it (mean in the middle, variance on the right)

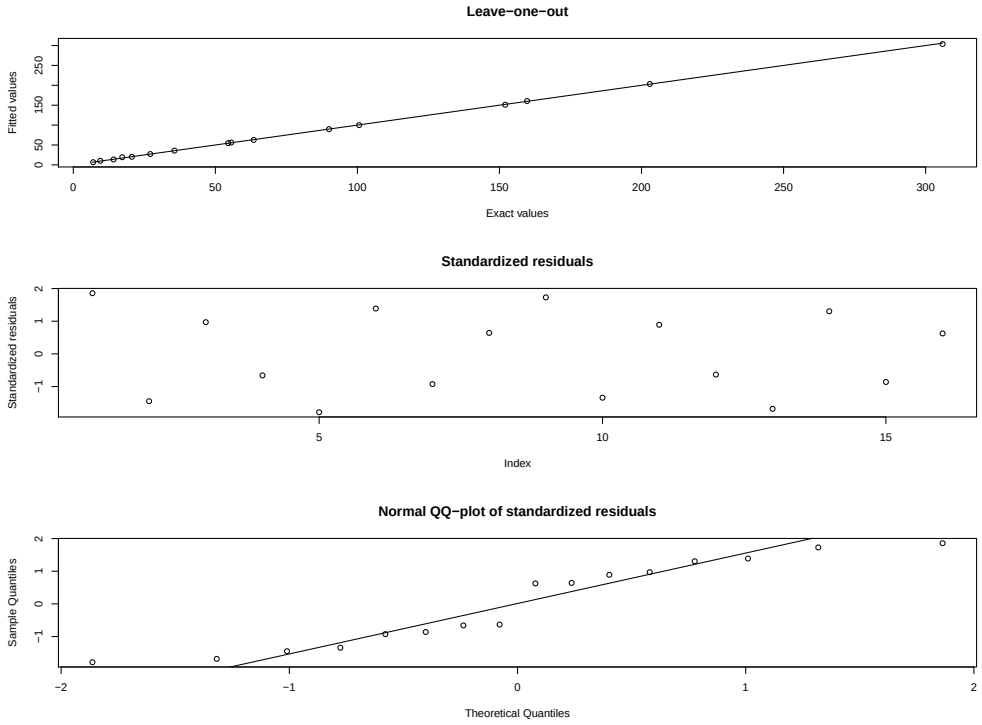

Figure 9: Leave-one-out cross-validation for the previous Kriging metamodel of the Branin-Hoo function

```
> n <- 80; d <- 6
> X <- randomLHS(n,d)
> X <- data.frame(X)
> y <- apply(X, 1, hartman6)
> mlog <- km(design=X, response=-log(-y))
> plot(mlog)
```

To go further, since the 6-dimensional Hartman function is cheap-to-evaluate, we study the performance of the model on a 250-point test LH sample generated at random. In practice,

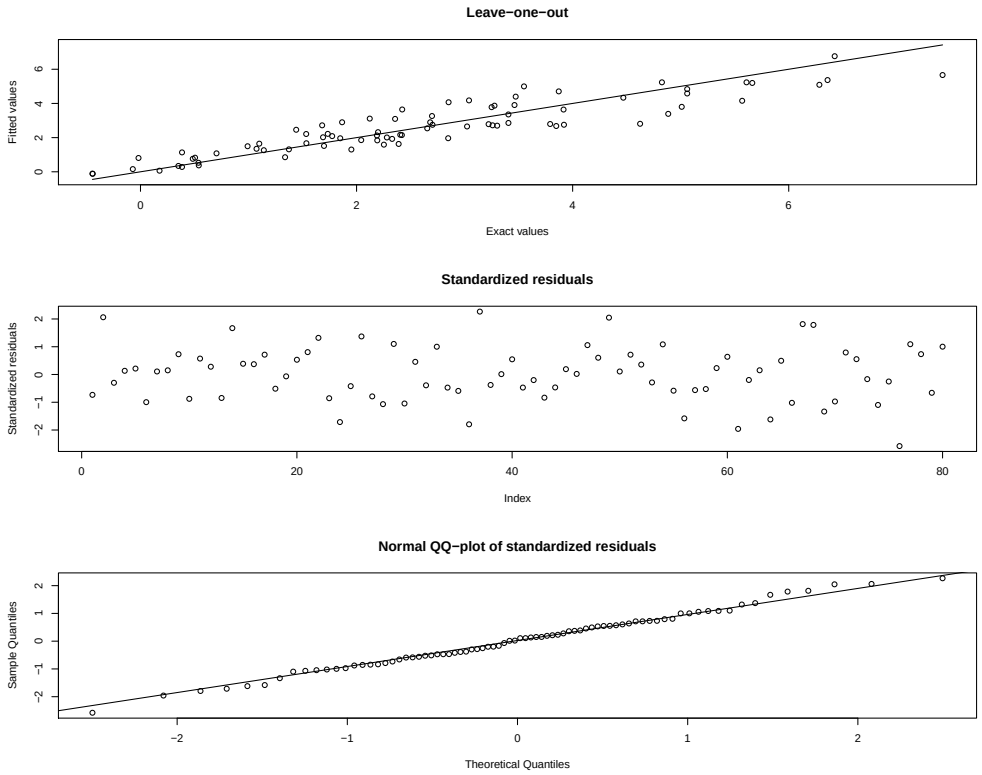

Figure 10: Leave-one-out validation for hartman6

of course, such a validation scheme would be intractable, and k-fold cross-validation would be a sensible substitute for it. Coming back to out-of-sample validation, we draw the predicted values versus the true ones (figure 11). We have also added the results that would be obtained with a trend (first order polynomial + interactions) or without the *log* transformation of the response. In this case, it seems clear that the log transformation is necessary for prediction. On the other hand, adding a trend does not result in obvious improvements.

```
> n.test <- 250
> X.test <- randomLHS(n.test, d)
> y.test <- apply(X.test, 1, hartman6)
> ylog.pred <- predict(mlog, newdata=X.test, type="UK")$mean
```

Performances of the estimation procedure: an empirical study

In order to study the performances of the estimation procedure implemented in `km`, we propose to perform estimation based on sample functions drawn from gaussian processes underlying prescribed Kriging models. In other words, we simulate sample functions corrspo kriging model which parameters are known. Then for each one of the sample functions (100 per Kriging model), we fit a Kriging model with `km`. We can then finally analyze the empirical distribution of these estimates, and make comparisons to the true ones (especially in terms of bias and variance).

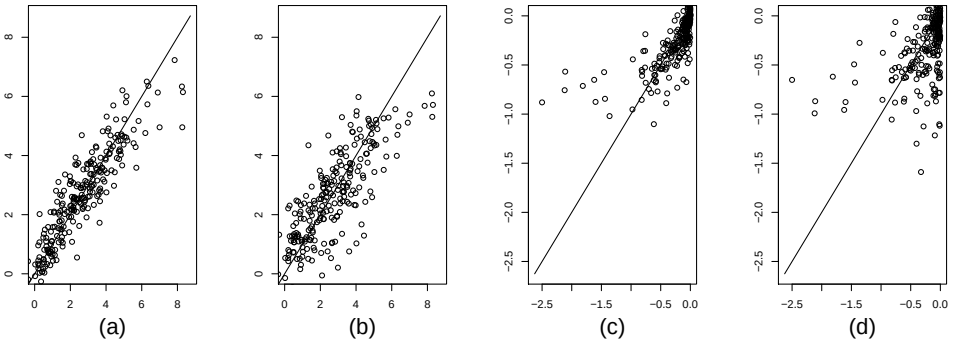

Figure 11: Out-of-sample validation for the 6-dimensional Hartman function on a 250-point test sample, with log transformation of the response (constant trend (a), or polynomial trend (b)) or without transformation (constant trend (c), polynomial trend (d)). The polynomial trend is a first order polynomial plus all 2nd order interactions.

In this paragraph, we show the results obtained in 3-dimensions and 10-dimensions. Following Loeppky, Sacks, and Welch (2009), we fix the number of runs n proportionaly to the problem dimension d. However, we have chosen $n = 15d$, instead of the informal rule "$n = 10d$", since - without any penalization of the likelihood - this seems to give more stable results in estimation. Then, we have used a maximin LH design, obtained with package **lhs**. Note that with that a LH design with such a few number of runs, only large enough values of the $\theta's$ can be estimated. Thus the $\theta's$ are chosen equally spaced from 0.3 to 0.7 in the 3-dimensional case, and equally spaced from 0.5 to 1.5 in the 10-dimensional one (in both cases, the domain is $[0, 1]^d$). The arguments of `km` are the default values, except in 10-dimensions where the upper bound was fixed to 3. In particular, the optimization method is BFGS. Better results can be obtained with the genetic algorithm, but will depend on its parameters, and thus are not shown here. The results are shown in figures 12 and 13.

4.4. The case of noisy observations

In the case where the observations assumed to be noisy, whether they steem from stochastic simulation (Fernex, Heulers, Jacquet, Miss, and Richet (2005)), from partially converged deterministic simulations (Forrester, Bressloff, and Keane (2006a)), or from real experiments, it is crucial to quantify the corresponding noise variance values and to take them into account within Kriging metamodeling.

Here we propose a basic one-dimensional example for all cases of Kriging with nugget effect, Kriging with homogeneous noise, and Kriging with heterogeneous noise (in reverse order here). Note that the results shown are directly transposable i) to multivariate cases, ii) involving Simple as well as Universal Kriging models.

```
> fundet <- function(x){
+ return((sin(10*x)/(1+x)+2*cos(5*x)*x^3+0.841)/1.6)}
> theta <- 1/sqrt(30)
> covtype <- "matern5_2"
> n <- 7
```

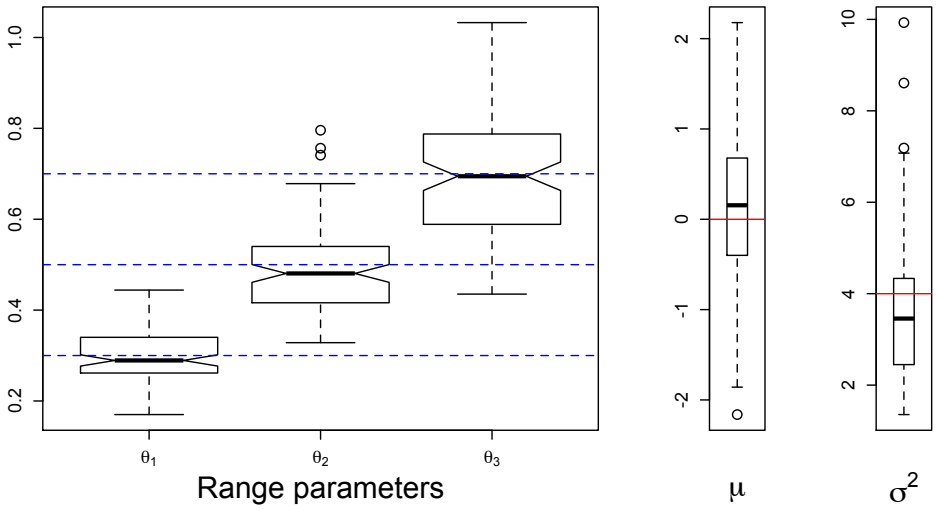

Figure 12: Empirical distribution of the parameter estimates. 3 dimensional case.

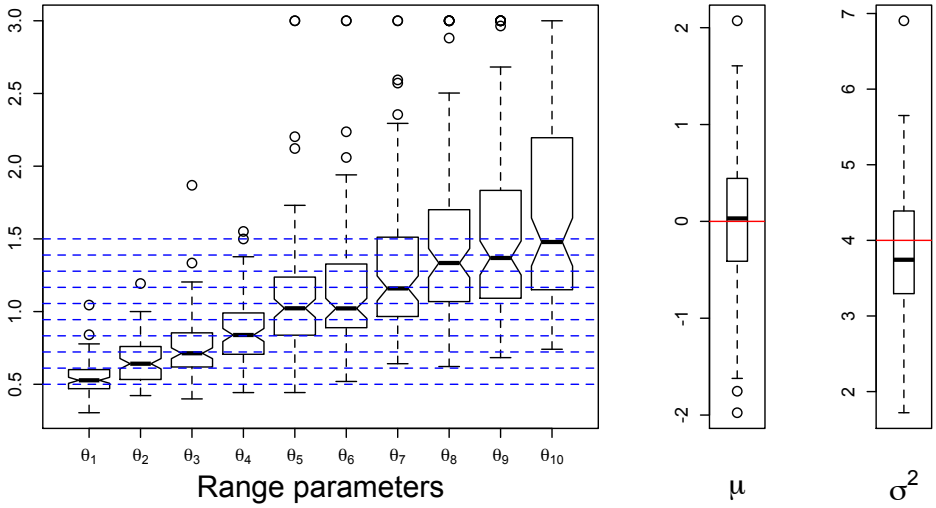

Figure 13: Empirical distribution of the parameter estimates. 10 dimensional case

```
> x <- seq(0,1, length=n)
> t <- seq(0,1,by=0.01)
> t <- sort(c(t,x))
> repart <- c(150,30,70,100,10,300,40)
> noise.var <- 4/repart
> z <- fundet(x); y <- z + sqrt(noise.var)*rnorm(length(x))
> model <- km(y~1, design=data.frame(x=x), response=data.frame(y=y), coef.trend=0,
+ covtype=covtype, coef.cov=theta, coef.var=1, noise.var=noise.var)
> p <- predict.km(model, newdata=t, type="SK")
```

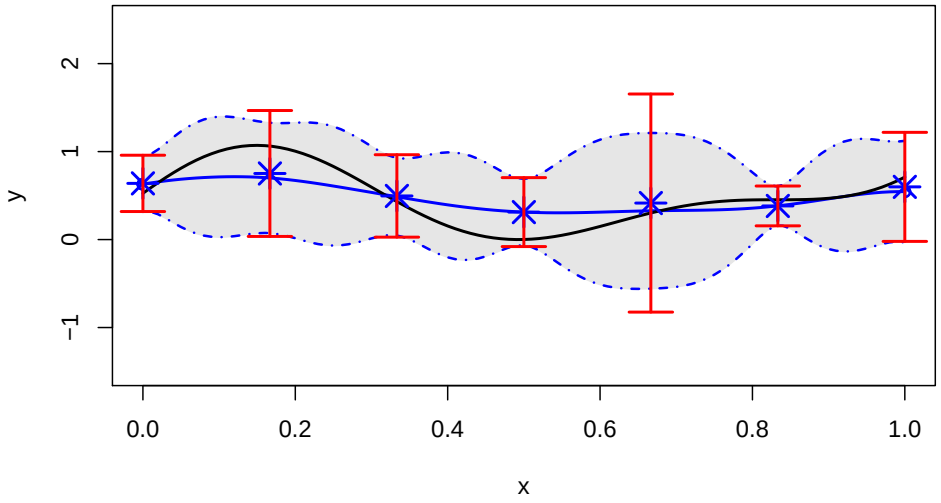

Figure 14: Kriging with heterogeneously noisy observations

The way the vectors `repart` and `noise.var` are coded correspond to the situation of a Monte-Carlo simulator with a total budget of 700 samplings heterogeneoulsy spread among the 7 points, with a distribution given by `repart <- c(150,30,70,100,10,300,40)` and a unitary variance of 4. Figure 15 illustrates the case where the same total computational budget is homogeneously distributed between the 7 points (`repart <- rep(100,7)`).

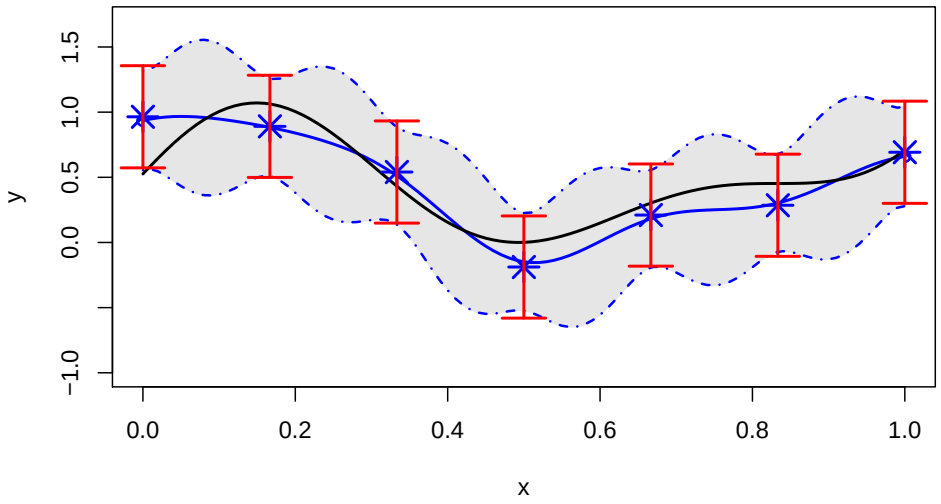

Figure 15: Kriging with homogeneously noisy observations

Finally, the same data is used in figure 16 to illustrate what a kind of Kriging model would have been obtained with a nugget instead of a homogeneous noise variance.

```
> model <- km(y~1, design=data.frame(x=x), response=data.frame(y=y),
+ coef.trend=0, covtype=covtype, coef.cov=theta, coef.var=1, nugget=4/100)
> p <- predict.km(model, newdata=sort(t), type="SK")
```

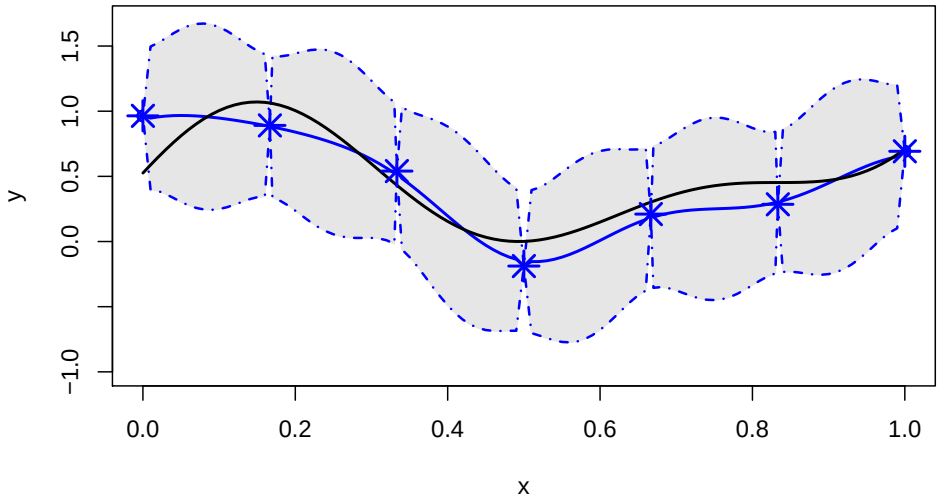

Figure 16: Kriging with homogeneously noisy observations

As announced in the Statistical Background section, the Kriging mean predictor with nugget effect is quite similar to the one with homogeneously noisy observations, up to its behaviour at the DoE. Indeed, one can see that the Kriging predictor interpolates the observations. As in the determenistic case, this goes with a zero Kriging variance at the corresponding points. Outside of the DoE, however, the Kriging variance has the same shape as in the analogue Kriging model with homogeneous noise, but with values translated by a factor of τ^2. This can be explained by the fact that the process Y doesn't have the same variance whether τ^2 stands for a nugget coefficient or for a noise variance.

4.5. Sensitivity analysis - Wraper to package Sensitivity

Sensitivity analysis (SA) is not implemented yet in this version of **DiceKriging**. However, it is easy to connect it to packages devoted to SA, and we suggest an efficient way of doing in this section. We refer to Sobol (1993) and Saltelli, Chan, and Scott (2000) for a presentation of SA. Let us illustrate the connection with the package **sensitivity**.

A toy example

Before considering an example suited to SA, first consider the Branin function. To perform SA with Branin, we first need to adapt its implementation to matrix-typed arguments.

```
> branin.mat <- function(X) apply(X,1,branin)
```

Then, assuming independent uniform distributions over $[-1, 1]$ for the inputs, the sobol indices can be computed using the function `fast99` [**sensitivity**], by typing (see help file of `fast99` for other examples):

```
> SA.branin <- fast99(model = branin.mat, factors = 2, n = 1000,
+ q = "qunif", q.arg = list(min = 0, max = 1))
```

Now, let us compute the SA of the kriging metamodel estimated from few runs of the Branin function again. After constructing the model (with a 16-point factorial design, as above),

```
> m.branin <- km(design=X, response=y)
```

we create a connection function based on `pred.km`, that returns only the kriging mean

```
> kriging.mean <- function(X, m) predict.km(m, X, "UK", se.compute=FALSE)$mean
```

to which we apply `fast99`:

```
> SA.metamodel <- fast99(model = kriging.mean, factors = 2, n = 1000,
+ q = "qunif", q.arg = list(min = 0, max = 1), m=m.branin)
```

The results can be printed, or drawn with a plot method taken from the package **sensitivity**. For the Branin function, the metamodel is precise, so the Sobol indices (main and total effects) calculated with the metamodel are very close to the true ones (figure 17).

```
> par(mfrow=c(1,2))
> plot(SA.branin); plot(SA.metamodel)
```

A standard SA 8-dimensional example

Let us now take the 8-dimensional Sobol function implemented in **sensitivity**. A kriging metamodel is estimated with a 80-point random LHS (generated by `randomLHS`[**lhs**]).

```
> n <- 80; d <- 8
> X <- randomLHS(n, d)
> X <- data.frame(X)
> y <- sobol.fun(X)
> m.sobol <- km(design=X, response=y)
```

The SA are computed as above:

```
> SA.metamodel <- fast99(model = kriging.mean, factors = d, n = 1000,
+ q = "qunif", q.arg = list(min = 0, max = 1), m=m.sobol)
> SA.sobol.fun <- fast99(model = sobol.fun, factors = d, n = 1000,
+ q = "qunif", q.arg = list(min = 0, max = 1))
```

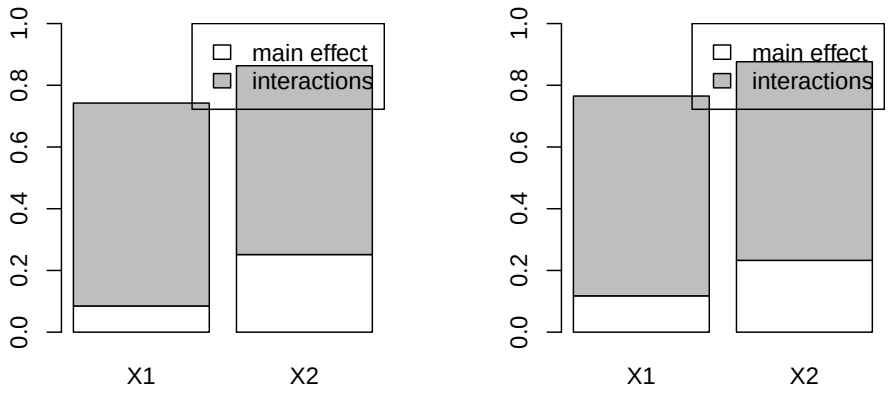

Figure 17: Sensitivity analysis for the Branin function (left) and a kriging metamodel of it (right)

Finally, the results are drawn on figure 18. The main characteristics are visible with the metamodel, but the Sobol indices relative to x_3 and x_4 are too small and may *not* be well captured, depending on the initial design.

```
> par(mfrow=c(1,2))
> plot(SA.sobol.fun); plot(SA.metamodel)
```

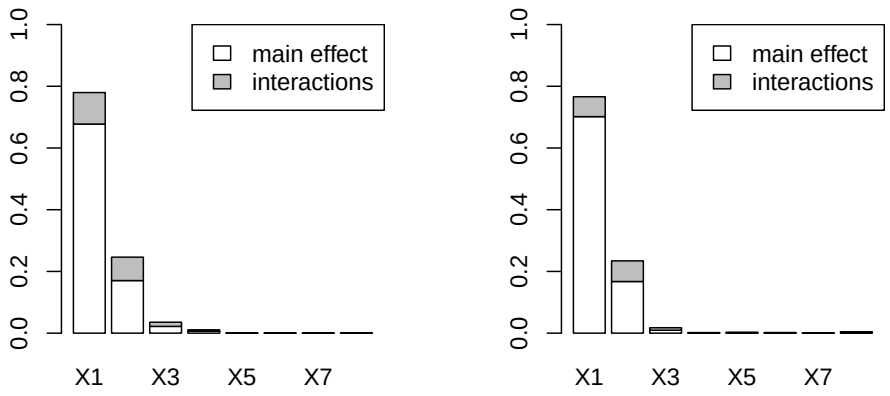

Figure 18: Sensitivity analysis for the Sobol function (left) and a kriging metamodel (right).

4.6. Known problems and their solutions

This section is is aimed at giving some illustrations to section 3.4, that should be read before.

Non invertibility of covariance matrices. Let us first consider again the example of the last section, but increase the size of the design of experiments.

```
> X <- expand.grid(seq(0,1,length=10), seq(0,1,length=10))
> X <- data.frame(X); names(X)<-c("x1", "x2")
> y <- branin(X)
> t <- try(km(design=X, response=y, covtype="gauss"))
> cat(t)

Error in chol.default(R) :
  the leading minor of order 40 is not positive definite
```

An error message indicates that the covariance matrix could not be inverted. To overcome this difficulty, one can choose the diagonal inflation method:

```
> km(design=X, response=y, covtype="gauss", nugget=1e-8*var(y))
```

or replace the Gaussian covariance kernel by the (default) Matérn kernel ($\nu = 5/2$):

```
> km(design=X, response=y)
```

Identifiability issues and flat estimated models. To illustrate the potential vitues of penalizing, consider the sine function proposed by Li and Sudjianto (2005), and compare the estimation results obtained with (only) 6 design points by three procedures: MLE, PMLE with SCAD penalty function, and MLE constrained by $\theta \geq \theta_{min}$ (figure 19). The Gaussian covariance kernel is used, with a small nugget effect. The usual MLE gives an unrealistic value of θ, estimated to 0.15 approx., which seems much too small in comparison to the distances between design points. On the other hand, both modified estimation procedures give realistic estimation results. However, a difficulty still remains in proposing a general method for choosing either the tuning parameter λ in PMLE or θ_{min}. In the present case, λ has been estimated by cross validation, and θ_{min} fixed to the mean distance between design points.

5. Examples with DiceOptim

5.1. Expected Improvement: 1D and 2D illustrations

As recalled in section 3, the EI criterion is at the heart of all Kriging-based optimization approaches considered in **DiceOptim**. Let us illustrate it on a first 1-dimensional toy example adapted from `qEI`'s help file. 5 points and corresponding observations are arbitrary chosen, and a Kriging metamodel with linear trend and Gaussian covariance is fitted to them.

```
> x <- c(0, 0.4, 0.6, 0.8, 1)
> y <- 10*c(-0.6, 0, -2, 0.5, 0.9)
> theta <- 0.1; sigma <- 10; trend <- 5*c(-2,1)
> model <- km(~x,design=data.frame(x=x), response=data.frame(y=y),
+ coef.trend=trend, covtype="gauss", coef.cov=theta, coef.var=sigma^2)
```

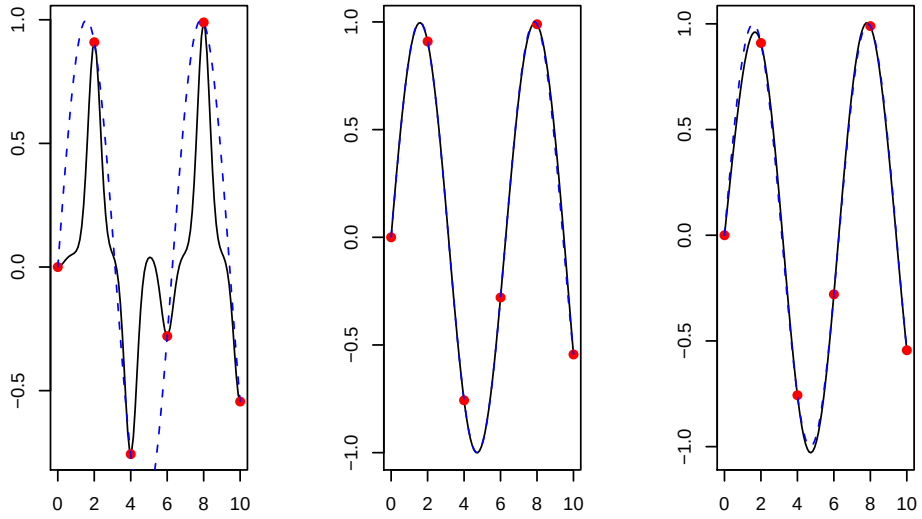

Figure 19: Comparison of three estimation methods for kriging approximation of the sine function, with 6 design points: MLE (left), PMLE with SCAD penalty (middle), MLE constrained by $\theta \geq 10/5$ (right). The bullets represent the design points, the solid line the Kriging mean and the dotted line the sine function.

The Kriging mean predictor, confidence intervals, and EI for $\mathbf{x}$ varying in $\in [0, 1]$ are then calculated and represented using the `predict` and `EI` functions:

```
> t <- seq(from=0, to=1, by=0.005)
> p <- predict(model, newdata=t, type="UK")
> EI_values <- apply(as.data.frame(t), 1, EI, model, type="UK")
```

We can observe on Figure 20 that EI behaves as described in the Statistical Background section: it is multimodal, null at the already sampled locations, and positive everywhere else with a magnitude increasing with both the decreasing Kriging mean and the increasing Kriging variance. A call to `EI` at one of the design points results indeed in a null value:

```
> EI(x[3], model, type="UK")

[1] 0
```

The maximum of EI is reached here at a unique point between the two first design points, where the uncertainty is the highest. A numerical maximization of the `EI` function can be obtained by using the dedication function `max.EI`, with tunable rectangular search domain, starting point, and selected control parameters for the underlying `rgenoud` algorithm:

```
> x_star <- max_EI(model, lower=0, upper=1, parinit=0.5, control =
+ list(pop.size=10, max.generations=10, wait.generations=5, BFGSburnin=10))
```

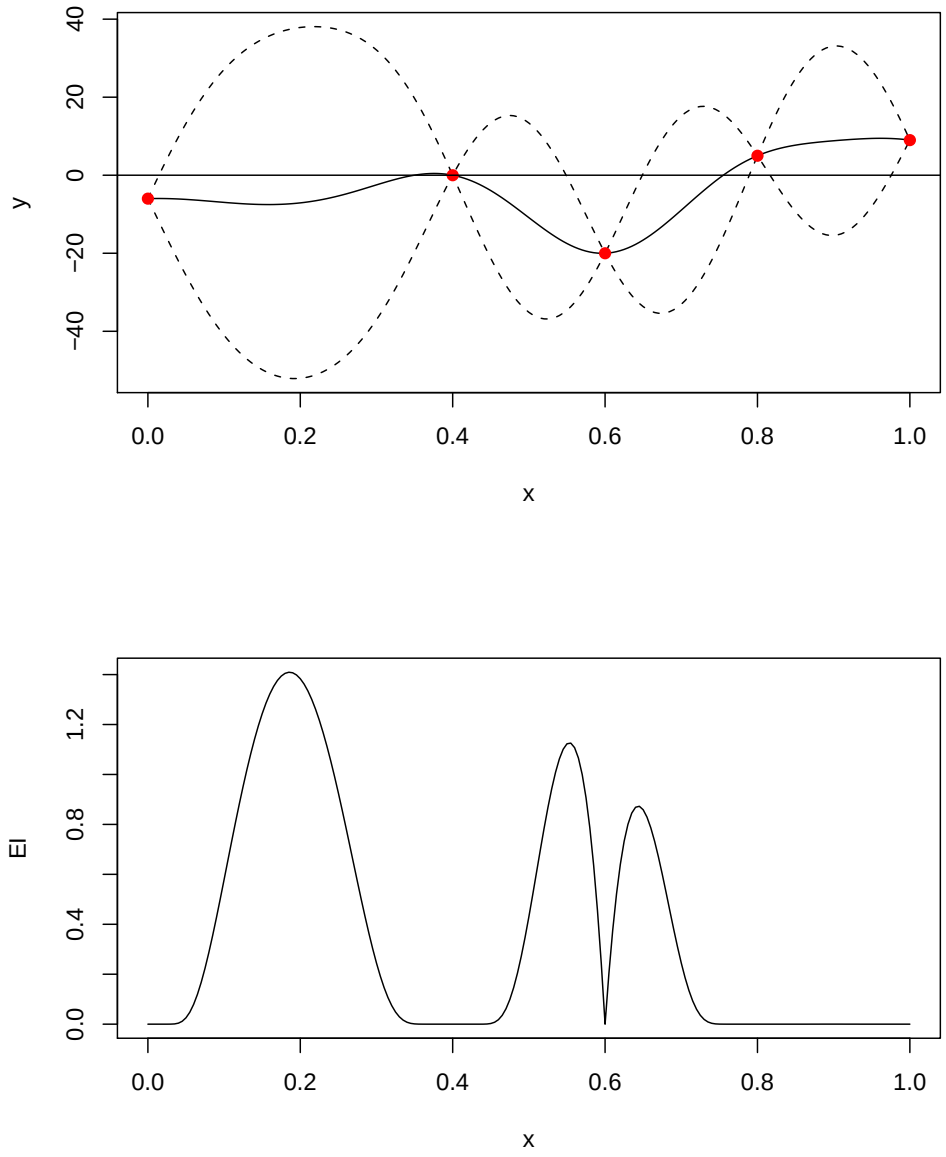

Figure 20: EI associated to a first 1-dimensional toy example

`max.EI` returns the optimal candidate point and corresponding EI value:

```
> print(x_star)

$par
              x
[1,] 0.1860575
```

```
$value
           EI
[1,] 1.409438
```

Let us now consider the 2-dimensional Branin function. This time, EI depends on a kriging model that needs to be estimated. In the sequel, the design is a random Latin Hypercube (LH) design of size 15. A kriging model is obtained with `km`, using the default values.

```
> d <- 2; n <- 15
> design <- randomLHS(n, d)
> design <- data.frame(design); names(design)<-c("x1", "x2")
> response.branin <- apply(design, 1, branin)
> fitted.model1 <- km(design=design, response=response.branin)
```

The corresponding EI is then computed over a grid:

```
> x.grid <- y.grid <- seq(0, 1, length=n.grid <- 25)
> design.grid <- expand.grid(x.grid, y.grid)
> EI.grid <- apply(design.grid, 1, EI, fitted.model1)
```

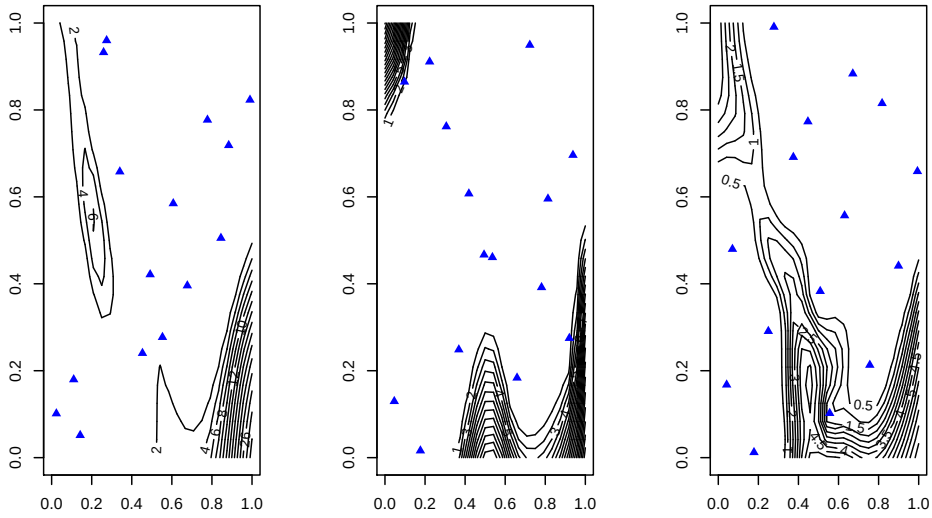

Figure 21: EI associated to Branin function

The results are represented on Figure 21 for three drawings of the LH design. In most cases, `EI` detects interesting regions when starting from 15 points, but the nature of the result may deeply differ depending on the inital design drawn. However, we will see in next section that the final optimization results are not very sensitive to the design choice, provided that enough points are sequentially added within EGO.

Finally, like in the 1-dimensional example, we observe that EI is multimodal. Thus a genetic algorithm is recommended for its optimization. To improve efficiency, the analytical gradient is implemented for the standard case (constant trend). One example of gradient field is represented in figure 22, obtained with a 3×3 factorial design (See help file of `EI.grad`).

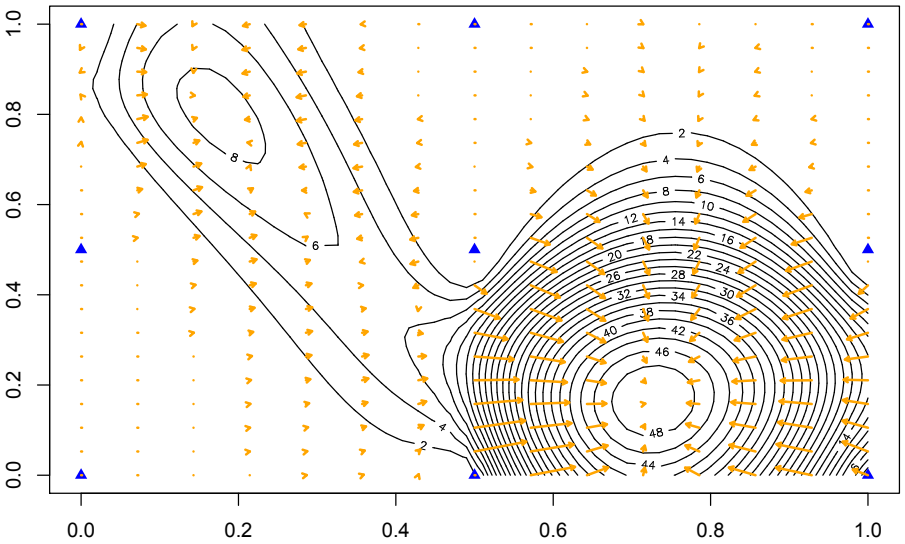

Figure 22: Gradient field of the EI associated to a Kriging metamodel of the Branin function

5.2. EGO illustrated on the Branin example

Now, let us apply the EGO algorithm to the Branin function. For the LH design used in last section, we run 10 steps of EGO by means of the `EGO.nsteps` function.

```
> # EGO n steps
> nsteps <- 10
> lower <- rep(0,d); upper <- rep(1,d)
> oEGO <- EGO.nsteps(model=fitted.model1, fun=branin, nsteps=nsteps,
+ lower, upper, control=list(pop.size=20, BFGSburnin=2))
```

The obtained sequence is shown on Figure 23 (left), as well as the EI contours corresponding to the last model (right). We observe that the 3 basins of global optima have been explored. Furthermore, the remaining EI after 10 steps is focused one these areas, showing that there isn't much interest to explore outside at this stage.

```
> par(mfrow=c(1,2))
> response.grid <- apply(design.grid, 1, branin)
> z.grid <- matrix(response.grid, n.grid, n.grid)
```

```
> contour(x.grid, y.grid, z.grid, 40)
> points(design[,1], design[,2], pch=17, col="blue")
> points(oEGO$par, pch=19, col="red")
> text(oEGO$par[,1], oEGO$par[,2], labels=1:nsteps, pos=3)
> EI.grid <- apply(design.grid, 1, EI, oEGO$lastmodel)
> z.grid <- matrix(EI.grid, n.grid, n.grid)
> contour(x.grid, y.grid, z.grid, 20)
> points(design[,1], design[,2], pch=17, col="blue")
> points(oEGO$par, pch=19, col="red")
```

Now, as pointed before, the whole EGO sequence depends on the initial LH design, and it is important to study the sensitivity of the results to the design choice. Thus, we have performed 10 steps of EGO with 100 random LH designs of size 15. Figure 24 represents the 3 points which are the closest to the 3 global optimizers of Branin function, for all 100 drawings. One may observe that the final result is nearly always satisfactory since all 3 regions are visited, and the optimum found is close to the true one.

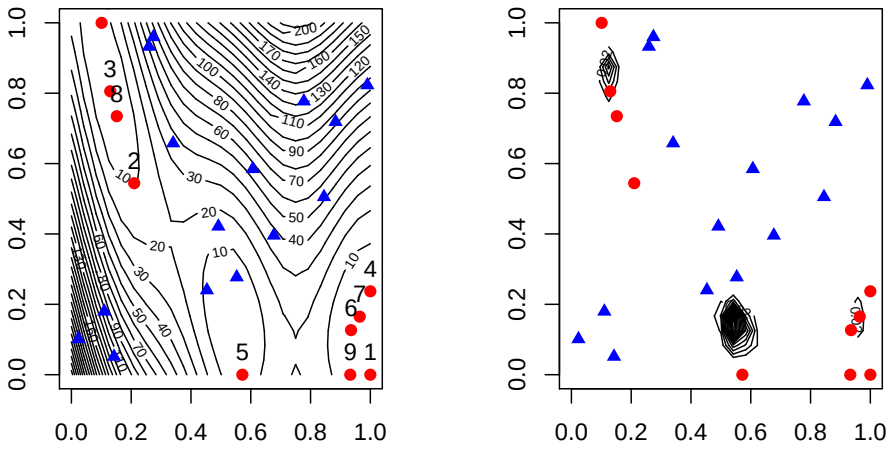

Figure 23: Sequence of visited points obtained 10 iterations of EGO applied to the Branin function (left), and EI contours corresponding to the last model (right)

5.3. Applications of EGO to the 6-dimensional Hartman function

We now come back to the 6-dimensional Hartman function previously considered, and build an optimization example upon it. Since it has been shown earlier that a logarithmic change of variable was necessary to obtain a well-bahaved Kriging metamodel, we choose to work here at first with suitably transformed data. The initial design chosen here is a 50-points design obtained by uniform sampling over $[0,1]^6$. For purpose of reproducibility, and since the variability is here greater than in 2 dimensions, one simulated design has been saved as `mydata` (the one of Ginsbourger (2009)), and is used all over the section.

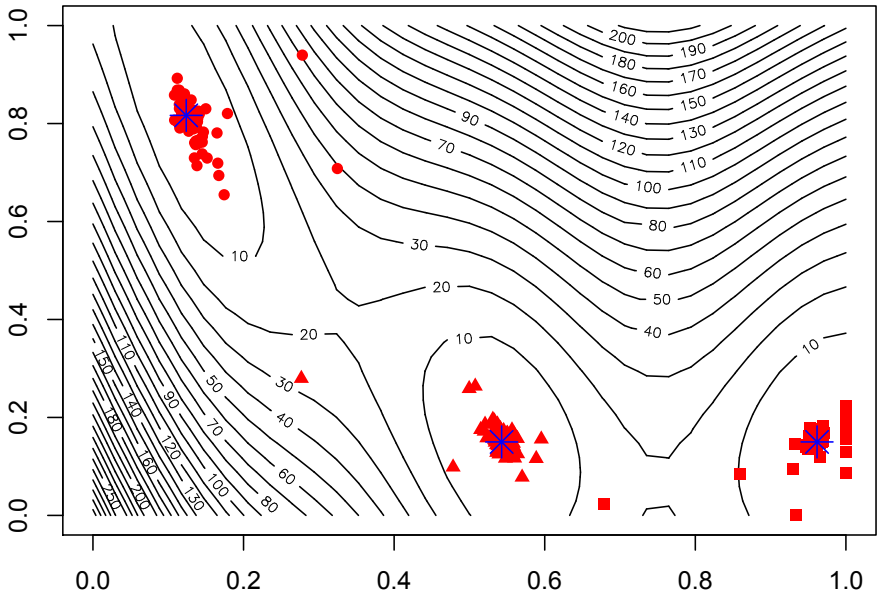

Figure 24: Sensitivity to the initial design of the EGO optimization results

```
> hartman6.log <- function(x) {-log(-hartman6(x))}
> data(mydata)
> X.total <- matrix(unlist(mydata), 50, 6)
> nb <- 50
> X <- X.total[1:nb, ]
> y <- apply(X, 1, hartman6.log)
```

An Ordinary Kriging metamodel with Matérn covariance ($\nu = \frac{5}{2}$) is fitted to the transformed data. Parameter estimation is performed by MLE, with the `rgenoud` optimizer. The control parameters are set such that the obtained results have a good level of reliability, which occurs a sligthly longer response time of `km` as with the default settings.

```
> hartman6.mm <- km(design=data.frame(X), response=y, control=list(pop.size=50,
+  max.generations=20, wait.generations=5, BFGSburnin=5), optim.method="gen")
```

We now apply 50 iterations of the EGO algorithm to the Hartman function, starting from the initial model above, based upon a 50-points design. The `EGO.nsteps` commando is commented to avoid long compilation times.

```
> nsteps <- 50
> # don't run
```

```
> # res.nsteps <- EGO.nsteps(model=hartman6.mm, fun=hartman6.log, nsteps=nsteps,
> # lower=rep(0,6), upper=rep(1,6), parinit=rep(0.5,6), control=list(pop.size=50,
> # max.generations=20, wait.generations=5, BFGSburnin=5), kmcontrol=NULL)
> #
> # To be compared with the current minimum of Harman6:
> hartman6.min <- -3.32
```

Starting from a 50*-point design*

As shown on Figure 25, EGO converges here to the actual global minimum (-3.32) of the Hartman function in less than 20 iterations when using a 50-points initial design.

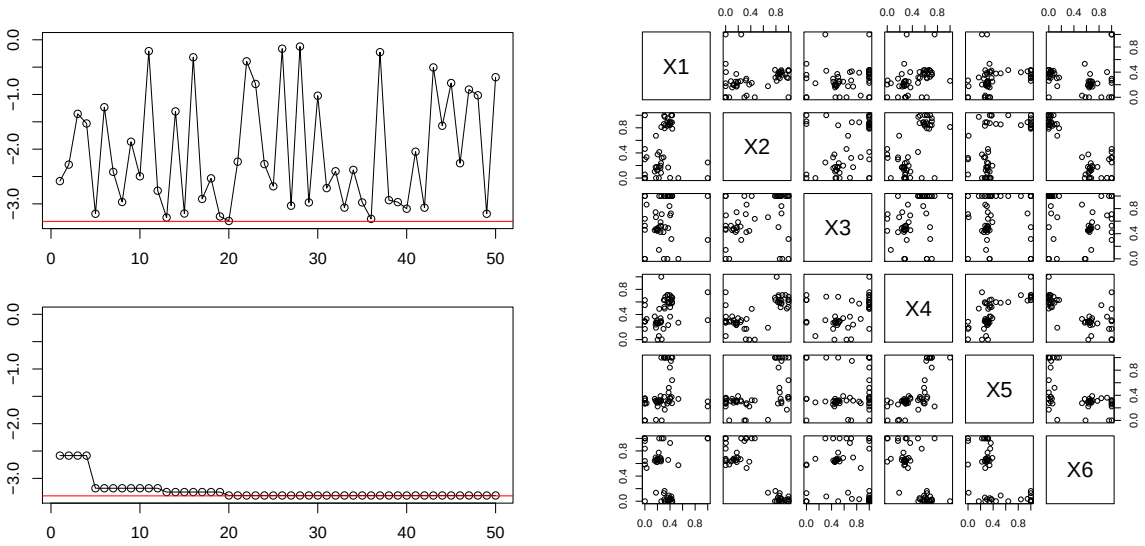

Figure 25: Optimization of the Hartman function with 50 initial points. Left: minimum value (top: current, bottom: cumulative). Right: pair plot of visited points during EGO.

Starting from a 10*-point design*

One of the crucial questions when using metamodel-based optimization algorithms with a severely restricted budget is the trade-off bewteen initial design points and additional points obtained during the algorithm itself. Answering this questionin a general way is of course out of scope here, but the following complementary experiment might contribute to motivate and illustrate the potential computational benefits of addressing this trade-off well.

Figure **?** shows the slower convergence of EGO when starting with a 10-points design. It is very interesting, however, to notice that the global expense in terms of total number of points is about twice smaller in the second case. Investigating the right trade-off in a generic framework seems a valuable research perspective.

Comparaison to the results obtained without a change of variable

To finish with examples concerning the regular EGO algorithm, let us try to apply the algo-

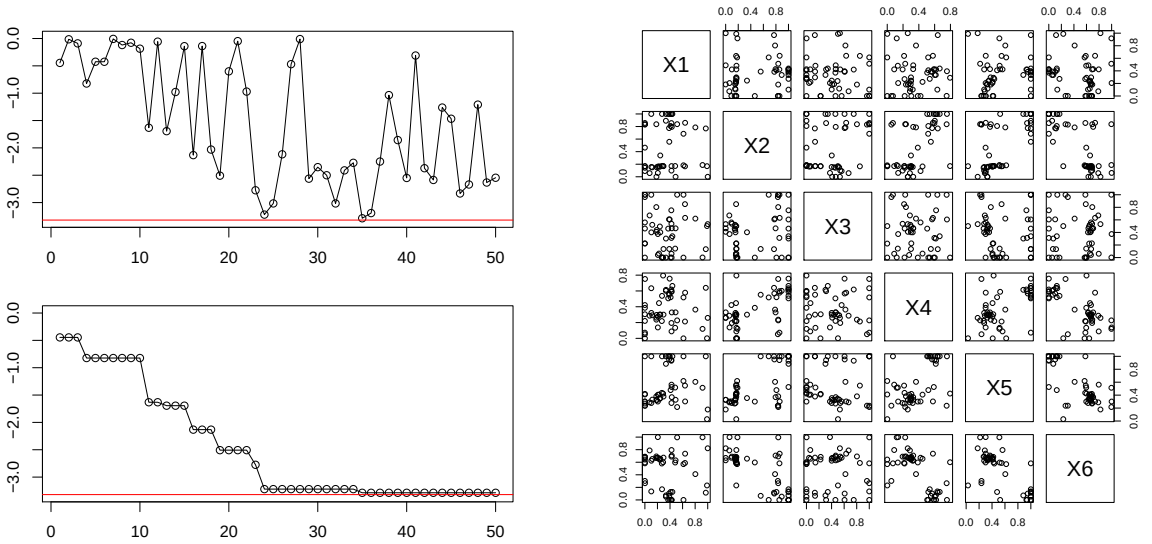

Figure 26: Optimization of the Hartman function with 10 initial points. Left: minimum value (top: current, bottom: cumulative). Right: pair plot of visited points during EGO.

rithm to the Hartman function, without any change of variables.

Figure 27 represents the results obtained with both previously considered initial designs.

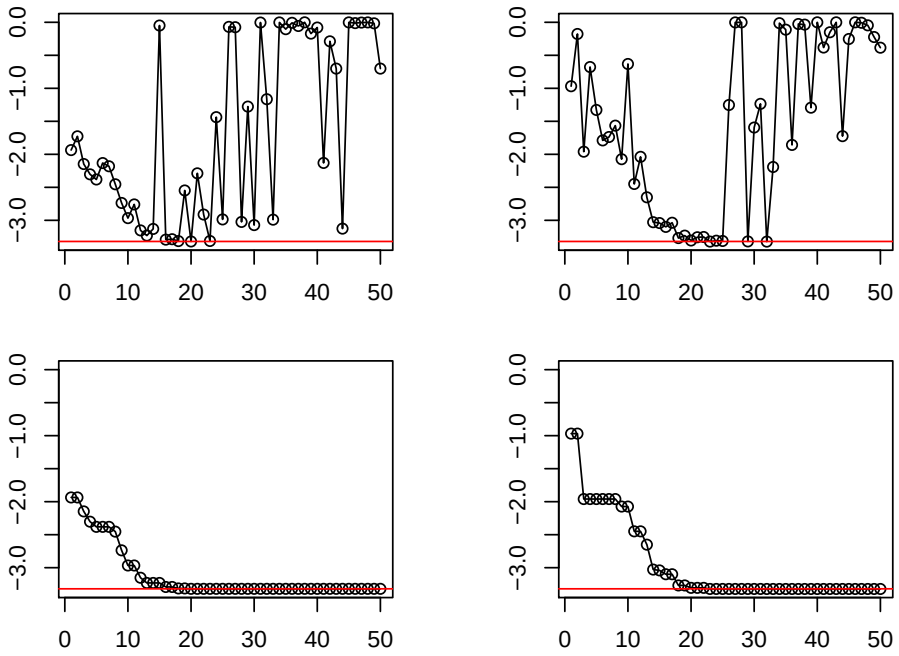

Figure 27: Hartman 6D. Without a change of variable. Left: 50 initial points, Right: 10 initial points.

Rather surprisingly, the obtained results are not worse than with an adapted transform, and even a bit better in the case of the 10-points initial design. This illustrates the robustness of EGO to some kinds of model misspecification, such non-gaussanity. To our knowledge, however, such algorithms may be very sensitive to misspecifications of the covariance kernel, in particular when it comes to correlation length parameters.

5.4. Parallelizations of EI and EGO

Distributing metamodel-based algorithms on several processors has become a contemporary industrial challenge since taking advantage of existing parallel facilities is a key to increase optimization performances when the time budget severely restricted.

q-points EI

DiceOptim includes a `qEI` function dedicated to the Monte Carlo estimation of the multipoints EI criterion. Here we come back to the first 1-dimensional example considered in the present section, and give an illustration of both 1-point and 2-points EI criteria estimations.

```
> candidate.design <- seq(0,1,,101)
> res <- qEI(newdata=candidate.design, model=model, type=type,
+ MC.samples=10000, return.I=TRUE)
> EI_estimated <- colMeans(res$I)
> EI_analytical <- apply(as.data.frame(candidate.design), 1, EI, model, type="UK")
> two_points_EI <- matrix(0,ncol=length(candidate.design),
+ nrow=length(candidate.design))
> for(i in seq(1,length(candidate.design)) ){
+ for(j in seq(i,length(candidate.design)) ){
+ qI <- pmax(res$I[,i],res$I[,j])
+ two_points_EI[i,j] <- mean(qI)
+ two_points_EI[j,i] <- two_points_EI[i,j]}}
```

The results shown on Figure 28 illustrate the adequacy between the Monte Carlo estimation of EI by means of `qEI` and the analytical formula implemented in `EI`.

The 2-points EI is represented on Figure 29, where it appears that sampling at the two highest bumps is likely to offer the best joint performance, i.e. in terms of multipoints expected improvement. More on the two-points EI criterion can be found in Ginsbourger *et al.* (2010).

Iterated Constant Liar test with Branin

Let us finish this section with an illustration of the Constant Liar algorithm, an approximate multipoints EI maximizer, applied to the Branin function previously used in several examples.

```
> nsteps <- 3
> npoints <- 8
> lower <- rep(0,d); upper <- rep(1,d)
> oEGOparallel1 <- CL.nsteps(model=fitted.model1, fun=branin, npoints=npoints,
+ nsteps=nsteps, lower, upper, control=list(pop.size=20, BFGSburnin=2))
```

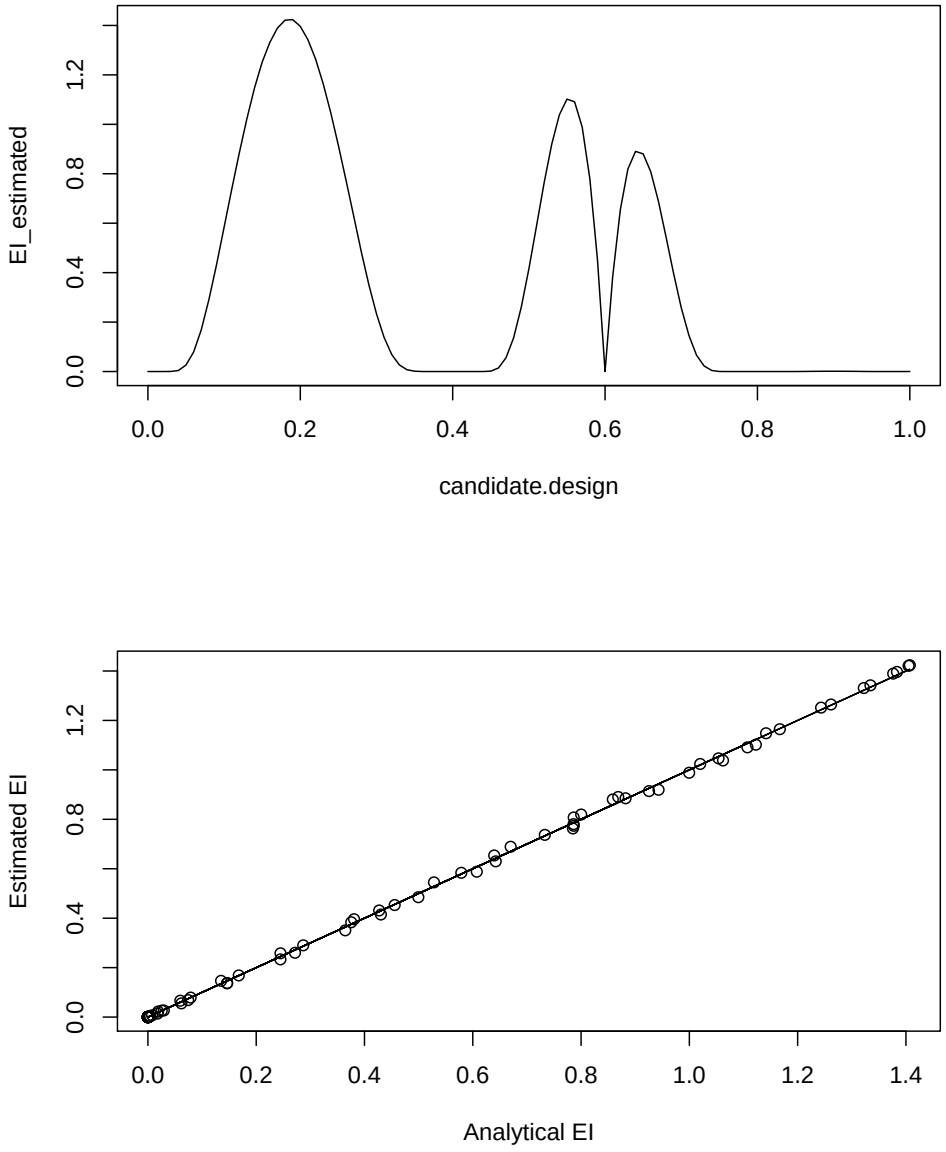

Figure 28: 1-point EI estimated by Monte Carlo and comparison to the previous results obtained with the analytical formula.

Starting from the initial LH design drawn in the previous Branin example, 3 iterations of Constant Liar with 8 points are applied sequentially by means of the `CL.nsteps` function (See results on Figure 30). Basically, each iteration of `CL.nsteps` consists in the creation of a 8-point design relying on the Constant Liar strategy, on the evaluation of the objective function at these points, and on the update of the Kriging model with the latter actual values.

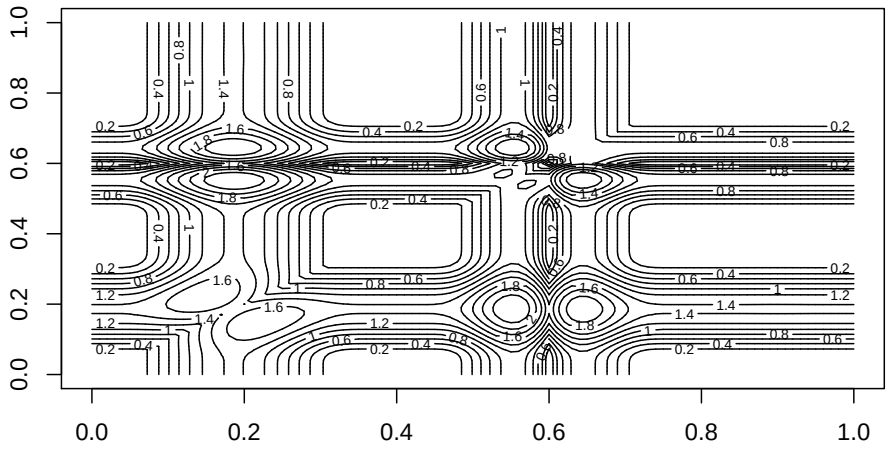

Figure 29: 2-points EI estimated by Monte Carlo

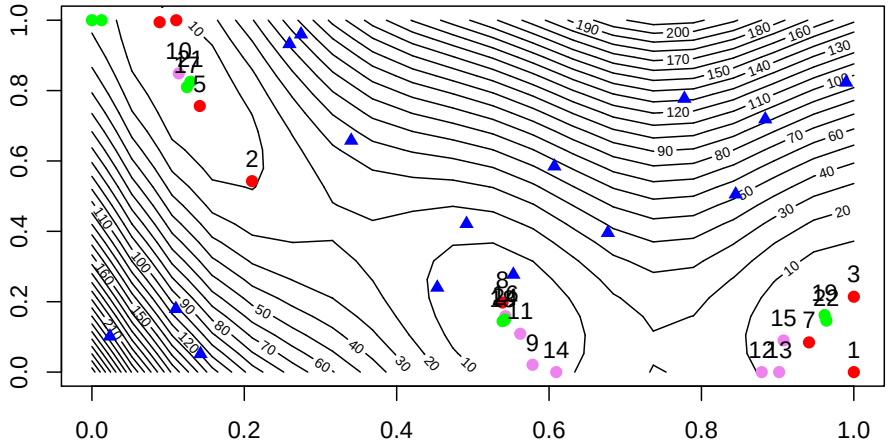

Figure 30: Parallelized EGO for the Branin function. 3 iterations of Constant Liar with 8 parallel searches.

Acknowledgments

This work was conducted within the frame of the DICE (Deep Inside Computer Experiments) Consortium between ARMINES, Renault, EDF, IRSN, ONERA and TOTAL S.A. (http://www.dice-consortium.fr/). The authors wish to thank Laurent Carraro, Delphine Dupuy and Céline Helbert for fruitful discussions about the structure of the code. They also thank

Gregory Six and Gilles Pujol for their advices on practical implementation issues, Khouloud Ghorbel for tests on sensitivity analysis, as well as the DICE members for useful feedbacks. Finally they gratefully acknowledge Yann Richet (IRSN) for numerous discussions concerning the user-friendliness of these packages.

A. Expressions of likelihoods and analytical gradients

The computations of likelihoods, concentrated likelihoods and analytical gradients are based mostly on Park and Baek (2001). They have been adapted for the kriging models dealing with noisy observations. Beware that the formulas below must not be used directly for implementation. For numerical stability and speed considerations, one should use the Cholesky decomposition of the covariance matrix and more generally avoid directly inverting matrices. In **DiceKriging**, we have used the efficient algorithm presented in Park and Baek (2001). See also the implementations details given in Rasmussen and Williams (2006).

The vector of observations $\mathbf{y}$ is normal with mean $\mathbf{F}\boldsymbol{\beta}$ and covariance $\mathbf{C}$. Thus, with notations of section **??**, the likelihood is:

$$L = \frac{1}{(2\pi)^{n/2}\,|\mathbf{C}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})'\mathbf{C}^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})\right) \tag{15}$$

The matrix $\mathbf{C}$ depends at least on the covariance parameters $\boldsymbol{\theta}, \sigma^2$.

A.1. Kriging model for noise-free observations

In this case we have $\mathbf{C} = \sigma^2\mathbf{R}$, where $\mathbf{R}$ depends only on $\boldsymbol{\theta}$, and $L = L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$. Writing the first order conditions results in analytical expressions for $\boldsymbol{\beta}$ and σ^2 as functions of $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{R}^{-1}\mathbf{y} \qquad \widehat{\sigma}^2 = \frac{1}{n}(\mathbf{y}-\mathbf{F}\widehat{\boldsymbol{\beta}})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{F}\widehat{\boldsymbol{\beta}})$$

Therefore maximizing the likelihood (15) is equivalent to maximizing over $\boldsymbol{\theta}$ only the "concentrated" log-likelihood obtained by plugging in the expressions of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$:

$$-2\log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \boldsymbol{\theta}) = n\log(2\pi) + n\log\widehat{\sigma}^2 + \log|\mathbf{R}| + n$$

The expression of the analytical gradient is the following:

$$-2\frac{\partial \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \boldsymbol{\theta})}{\partial\theta_k} = -(\mathbf{y}-\mathbf{F}\widehat{\boldsymbol{\beta}})'\mathbf{R}^{-1}\frac{\partial\mathbf{R}}{\partial\theta_k}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{F}\widehat{\boldsymbol{\beta}})/\widehat{\sigma}^2 + \mathrm{tr}\left(\mathbf{R}^{-1}\frac{\partial\mathbf{R}}{\partial\theta_k}\right)$$

A.2. Kriging model for noisy observations (unknown homogeneous noise)

In this case we have $\mathbf{C} = \sigma^2\mathbf{R} + \tau^2\mathbf{I}_n$. The likelihood $L = L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2)$ also depends on the new nugget parameter τ^2, but it is possible to reduce the optimization dimensionality. Indeed, define:

- $v = \sigma^2 + \tau^2$, the total variance
- $\alpha = \frac{\sigma^2}{\sigma^2+\tau^2}$, the proportion of variance explained by $Z(.)$

We can rewrite $\mathbf{C} = v\mathbf{R}_\alpha$ with $\mathbf{R}_\alpha = \alpha\mathbf{R} + (1-\alpha)\mathbf{I}_n$. Note that $\mathbf{R}_\alpha$ is also symmetric positive-definite since $\alpha \in [0,1]$. Then writing the first order conditions results in analytical expressions for $\boldsymbol{\beta}$ and v as functions of $\boldsymbol{\theta}$ and τ^2:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{R}_\alpha^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{R}_\alpha^{-1}\mathbf{y} \qquad \widehat{v} = \frac{1}{n}(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})'\mathbf{R}_\alpha^{-1}(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})$$

The concentrated log-likelihood depends only on $\boldsymbol{\theta}$ and α:

$$-2\log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}, \widehat{v}, \boldsymbol{\theta}, \boldsymbol{\alpha}) = n\log(2\pi) + n\log\widehat{v} + \log|\mathbf{R}_\alpha| + n$$

The fact that α is bounded ($\alpha \in [0,1]$) is convenient for optimization. The derivatives with respect to the $\boldsymbol{\theta}_k$'s and α are given by:

$$-2\frac{\partial \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}, \widehat{v}, \boldsymbol{\theta}, \boldsymbol{\alpha})}{\partial \bullet} = -(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})'\mathbf{R}_\alpha^{-1}\frac{\partial \mathbf{R}_\alpha}{\partial \bullet}\mathbf{R}_\alpha^{-1}(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})/\widehat{v} + \mathrm{tr}\left(\mathbf{R}_\alpha^{-1}\frac{\partial \mathbf{R}_\alpha}{\partial \bullet}\right)$$

with $\frac{\partial \mathbf{R}_\alpha}{\partial \theta_k} = \alpha\frac{\partial \mathbf{R}}{\partial \theta_k}$ and $\frac{\partial \mathbf{R}_\alpha}{\partial \alpha} = \mathbf{R} - \mathbf{I}_n$.

A.3. Kriging model for noisy observations (known noise)

The likelihood $L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ takes the form above (15) but with $\mathbf{C} = \sigma^2\mathbf{R} + \mathrm{diag}(\tau_1^2, \ldots, \tau_n^2)$. Writing the first order conditions results in analytical expression for $\boldsymbol{\beta}$ only:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{C}^{-1}\mathbf{y}$$

The concentrated log-likelihood depends both on $\boldsymbol{\theta}$ and σ^2:

$$-2\log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}, \sigma^2, \boldsymbol{\theta}) = n\log(2\pi) + \log|\mathbf{C}| + (\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})'\mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})$$

The derivatives with respect to the $\boldsymbol{\theta}_k$'s and σ^2 are given by:

$$-2\frac{\partial \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}, \sigma^2, \boldsymbol{\theta})}{\partial \bullet} = -(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})'\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \bullet}\mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}) + \mathrm{tr}\left(\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \bullet}\right)$$

with $\frac{\partial \mathbf{C}}{\partial \theta_k} = \sigma^2\frac{\partial \mathbf{R}}{\partial \theta_k}$ and $\frac{\partial \mathbf{C}}{\partial \sigma^2} = \mathbf{R}$.

A.4. Case of known trend

When the vector $\boldsymbol{\beta}$ of trend parameters is known, one can check that all the aforementioned formula for (concentrated) likelihoods and derivatives still stand by replacing $\widehat{\boldsymbol{\beta}}$ by $\boldsymbol{\beta}$.

B. Analytical gradient of Expected Improvement

The aim of this section is to present the efficient algorithm developed for package **DiceOptim** to compute the analytical gradient of expected improvement (EI) in the most common case: noise-free observations with constant mean (Ordinary Kriging). The method adapts for a general linear trend but requires the implementation of the derivatives of all trend basis functions $\mathbf{f}_k$. First recall the EI expression:

$$\mathrm{EI}(\mathbf{x}) = (a - m(\mathbf{x})) \times \Phi(z(\mathbf{x})) + s(\mathbf{x}) \times \phi(z(\mathbf{x}))$$

where a is the current function minimum value, $m(\mathbf{x}) = m_{\text{UK}}(\mathbf{x})$ and $s^2(\mathbf{x}) = s^2_{\text{UK}}(\mathbf{x})$ are the prediction mean and variance for universal kriging, and $z(\mathbf{x}) = (a - m(\mathbf{x}))/s(\mathbf{x})$. By using the relations $\Phi' = \phi$ and $\phi'(t) = -t\phi(t)$, the gradient of $EI(\mathbf{x})$ reduces to:

$$\nabla \text{EI}(\mathbf{x}) = -\nabla m(\mathbf{x}) \times \Phi(z(\mathbf{x})) + \nabla s(\mathbf{x}) \times \phi(z(\mathbf{x})) \tag{16}$$

Denote $\hat{\mu} = \frac{\mathbb{1}'\mathbf{C}^{-1}\mathbf{y}}{\mathbb{1}'\mathbf{C}^{-1}\mathbb{1}}$. Then for a constant trend we have:

$$m(\mathbf{x}) = \widehat{\mu} + \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}(\mathbf{y} - \mathbb{1}\widehat{\mu}) \qquad s^2(\mathbf{x}) = \widehat{\sigma}^2 - \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}) + \frac{(1 - \mathbb{1}'\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}))^2}{\mathbb{1}'\mathbf{C}^{-1}\mathbb{1}}$$

From which, using that $\nabla s^2(\mathbf{x}) = 2s(\mathbf{x})\nabla s(\mathbf{x})$, we deduce:

$$\begin{aligned} \nabla m(\mathbf{x}) &= \nabla \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}(\mathbf{y} - \mathbb{1}\widehat{\mu}) \\ \nabla s(\mathbf{x}) &= -\frac{1}{s(\mathbf{x})}\left(\nabla \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}) + \frac{(1 - \mathbb{1}'\mathbf{C}^{-1}\mathbf{c}(\mathbf{x}))\nabla \mathbf{c}(\mathbf{x})'\mathbf{C}^{-1}\mathbb{1}}{\mathbb{1}'\mathbf{C}^{-1}\mathbb{1}}\right) \end{aligned}$$

To compute these expressions efficiently, we use the Cholesky decomposition of the covariance matrix and the resulting auxiliary variables $\mathbf{z}, \mathbf{M}, \mathbf{u}$ defined in appendix C.1. As $\mathbf{F} = \mathbb{1}$, $\mathbf{M}$ is a $n \times 1$ vector, and is renamed $\mathbf{u}$ in this section. In the same way, we introduce the $n \times d$ matrix $\mathbf{W} = (\mathbf{T}')^{-1}(\nabla \mathbf{c}(\mathbf{x})')'$. Then we can rewrite $\nabla m(\mathbf{x})$ and $\nabla s(\mathbf{x})$ in the concise form

$$\nabla m(\mathbf{x}) = \mathbf{W}'\mathbf{z} \qquad \nabla s(\mathbf{x}) = -\frac{1}{s(\mathbf{x})}\left(\mathbf{W}'\mathbf{v} + \frac{(1 - \mathbf{v}'\mathbf{u})(\mathbf{W}'\mathbf{u})}{\mathbf{u}'\mathbf{u}}\right)$$

and $\nabla \text{EI}(\mathbf{x})$ is obtained with formula (16).

Computational cost We now indicate the order of the marginal computational cost, knowing the results of past computations. In particular we assume that a `km` object was first created, and thus the auxiliary variables $\mathbf{T}, \mathbf{z}, \mathbf{u}$ have already been computed. In addition, $\nabla \text{EI}(\mathbf{x})$ is supposed to be computed after that $\text{EI}(\mathbf{x})$ was evaluated, as it is the case during the optimization procedure. Finally we ignore the terms of order n (assuming that $n \gg d$ and $n \gg p$). Then the complexity of $\text{EI}(\mathbf{x})$ is given by the kriging mean computation step, equal to n^2 (see appendix C.2). During this step, $s(\mathbf{x}), \Phi(z(\mathbf{x}))$ and $\phi(z(\mathbf{x}))$ are stored. Therefore, the complexity for $\nabla \text{EI}(\mathbf{x})$ is due to the computation of $\mathbf{W}$, which is done by solving d upper triangular systems of linear equations, and to some linear algebra. This requires dn^2 operations.

C. Implementation notes

C.1. Auxiliary variables

To prevent from numerical stabilities and avoid re-computations, four auxiliary variables are used in **DiceKriging** and **DiceOptim**, three of them were proposed by Park and Baek (2001) and the fourth was added for prediction:

- $\mathbf{T}$, $n \times n$ matrix: the upper triangular matrix obtained in the Cholesky decomposition of the (positive definite) covariance matrix. Thus we have: $\mathbf{C} = \mathbf{T}'\mathbf{T}$

- $\mathbf{z} = (\mathbf{T}')^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})$, $n \times 1$ vector
- $\mathbf{M} = (\mathbf{T}')^{-1}\mathbf{F}$, $n \times p$ matrix
- $\mathbf{v} = (\mathbf{T}')^{-1}\mathbf{c}(\mathbf{x})$, $n \times 1$ vector

In the case where parameters are estimated, the expressions for $\mathbf{z}$ and $\mathbf{v}$ are modified by replacing the true parameters by their ML estimate, and could be hatted: $\widehat{\mathbf{z}} = (\widehat{\mathbf{T}}')^{-1}(\mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}})$, $\widehat{\mathbf{v}} = (\widehat{\mathbf{T}}')^{-1}\widehat{\mathbf{c}}(\mathbf{x})$. Actually for sake of simplicity, no distinction is made.

C.2. Formulas for prediction

In this paragraph, we give the formulas used in **DiceKriging** to implement the kriging mean and variance for simple and universal kriging. Some computation economy and numerical stability can be obtained by re-writing the formulas with the auxiliary variables already computed (except $\mathbf{v}$) in creating a `km` object. Firstly, with notations of section C.1, the formulas of simple kriging become:

$$m_{\mathrm{SK}}(\mathbf{x}) = \mathbf{f}(\mathbf{x})'\boldsymbol{\beta} + \mathbf{v}'\mathbf{z} \qquad s^2_{\mathrm{SK}}(\mathbf{x}) = C(\mathbf{x},\mathbf{x}) - \|\mathbf{v}\|^2$$

where $\|.\|$ denotes the Euclidian distance. For universal kriging, expressions are similar but require the Cholesky decomposition of $\mathbf{F}'\mathbf{C}^{-1}\mathbf{F} = \mathbf{M}'\mathbf{M}$. Denote $\mathbf{T}_{\mathrm{M}}$ the corresponding $p \times p$ upper triangular matrix s.t. $\mathbf{T}_{\mathrm{M}}{}'\mathbf{T}_{\mathrm{M}} = \mathbf{M}$. Then we have:

$$m_{\mathrm{UK}}(\mathbf{x}) = \mathbf{f}(\mathbf{x})'\widehat{\boldsymbol{\beta}} + \mathbf{v}'\mathbf{z} \qquad s^2_{\mathrm{UK}}(\mathbf{x}) = C(\mathbf{x},\mathbf{x}) - \|\mathbf{v}\|^2 + \|(\mathbf{T}_{\mathrm{M}}{}')^{-1}(\mathbf{f}(\mathbf{x}) - \mathbf{M}'\mathbf{v})\|^2$$

In the last expression, $(\mathbf{T}_{\mathrm{M}}{}')^{-1}(\mathbf{f}(\mathbf{x}) - \mathbf{M}'\mathbf{v})$ is obtained by solving an upper triangular system of size $p \times p$.

Computational cost Ignoring terms of order n (assuming that $n \gg d$ and $n \gg p$), the complexity for kriging mean is given by the backsolving of $\mathbf{T}'\mathbf{v} = \mathbf{c}(\mathbf{x})$ to get $\mathbf{v}$, which is equal to n^2. The complexity for kriging variance is smaller: $2n$ for SK, $2n(p+1) + \left(\frac{p^3}{3} + p^2\right)$ for UK *[the additional non-negligible operations decompose as follows: $2np$ for $\mathbf{M}'\mathbf{v}$, $\frac{1}{3}p^3$ for Cholesky decomposition of $\mathbf{M}$, and p^2 to compute the matrix product $(\mathbf{T}_M{}')^{-1}(\mathbf{f}(\mathbf{x}) - \mathbf{M}'\mathbf{v})$]*.

C.3. Table of computational cost and memory size

In the table below, we give an estimate of the computational cost and memory size required for the most important procedures implemented in **DiceKriging** and **DiceOptim**. The computational cost is for a single procedure, and assumes that some stored variables need not be re-computed. For instance when running the `predict` method, a `km` object was first created, and the auxiliary variables of section C.1 are already stored. In addition, we assume that $n \gg d$ and $n \gg p$. The memory size represents here the order of magnitude of the quantity of numbers to be stored at the same time when running the procedure. For prediction, results are given as a function of m, the number of new data at which we want to predict, since this number is usually large. As for n, we assume that $m \gg d$ and $m \gg p$.

Procedure	**Complexity**	**Memory**	**Limiting step**
log-Likelihood	$\frac{1}{3}n^3 + \left(\frac{d+p}{2} + 2\right) n^2$	n^2	Cholesky dec. of $\mathbf{C}$
log-Likelihood gradient	$\frac{1}{3}n^3 + (6d+1)n^2$	n^2	Inverting $\mathbf{C}$ from $\mathbf{T}$
Kriging mean (SK, UK)	mn^2	$(n+m)n$	$(\mathbf{T}')^{-1}\mathbf{c}(\mathbf{x})$
Kriging variance (SK)	$2mn$	$(n+m)n$	
Kriging variance (UK)	$2m(p+1)n + m\left(\frac{p^3}{3} + p^2\right)$	$(n+m)n$	
EI	n^2	n^2	Kriging mean
EI gradient	dn^2	n^2	$(\mathbf{T}')^{-1}(\nabla\mathbf{c}(\mathbf{x})')'$

Table 4: Computational cost (complexity) and memory size of the most important procedures implemented in **DiceKriging** and **DiceOptim**. Recall that n is the number of design points, d the problem dimension, p the number of trend basis functions and m the number of new data where to predict.

This table can be useful to guess the difficulties linked to the increasing of the design size. For instance the log-likelihood complexity is approximately $\frac{1}{3}n^3$: multiplying by 2 the number of experiments results in multiplying by 8 the computational time, and by 4 the size of the covariance matrix involved.

Comments. 1. The complexity related to log-likelihoods is from Park and Baek (2001); The factors $\left(\frac{d+p}{2} + 2\right) n^2$ and $(6d+1)n^2$ are for Gaussian covariance: it can be larger for more complex covariances but will not modify the order of magnitude. For other procedures, see appendices B and C.2. 2. The memory size is no less than n^2, which corresponds to the storage of the $n \times n$ covariance matrix. When computing gradients, only the final derivatives are stored, and not the auxiliary matrix derivatives. The memory size for computing the kriging mean and variance can be large: this is because we have vectorized the computation of formulas described in C.2, to improve speed.

C.4. Speed

Some functions have been implemented in the C language. There was no need to do so with linear algebra R routines, since they call themselves Fortran or C routines. On the other hand, the computation of covariance matrices seems to be very slow in R, due to the presence of triple loops. They have been implemented in C (functions `covMatrix`, `covMat1Mat2`).

D. Interface considerations

In some rare contexts the simulator function can be turned into a R function using external facilities such as `.C` or `.Call`. Most likely, the simulation will be called from another computing environment such as java. This environment's event loop will then call R as well as perform some other tasks such as database management for complex input/output. The **Rserve** and the **Rsession** java packages are well suited for such a context: results of R computations can be temporarily stored as R objects. Interesting objects 'morally' correspond to two R classes: (spatial) *fitted model* and (optimization) *algorithm in progress*.

In such a context, R calls should be kept as simple as possible or be generated by the main environment. Challenging tasks are the management of the documentation and that of the formal arguments. A desirable feature for CE tools is that the analyst should not be required to have an extended knowledge in R, and thus be concerned only by high level functions.

The S4 mechanism is well suited for this context. The `km` class of **DiceKriging** and its classical methods are a first step to further S4 integration. An efficient `update` method could speed the computations in the future.

Parallelizable algorithms such as EGO can be viewed as an iterated couple of steps: *ask* for a new design point and *tell* a new response to the model/algorithm object, the two steps being repeated until convergence is reached. They can be implemented as calls to `ask` and `tell` methods (as in the **sensititvity** package) or as two variants of an `update` method.

References

Bartz-Beielstein T, Preuss M (2007). "Experimental research in evolutionary computation." In "Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation," .

Brochu E, Cora M, de Freitas N (2009). "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning." *Technical report*, Department of Computer Science, University of British Columbia.

Chambers J (2008). *Software for Data Analysis. Programming with R*. Springer.

Cressie N (1993). *Statistics for spatial data*. Wiley series in probability and mathematical statistics.

Emmerich M, Giannakoglou K, Naujoks B (2006). "Single-and Multiobjective Optimization Assisted by Gaussian Random Field Metamodels." *IEEE Transactions on Evolutionary Computation*, **10 (4)**.

Fan J (1997). "Comment on "Wavelets in Statistics: A Review by A. Antoniadis"." *Italian Journal of Statistics*, **6**, 97–144.

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

Fernex F, Heulers L, Jacquet O, Miss J, Richet Y (2005). "The MORET 4B Monte Carlo code - New features to treat complex criticality systems." In "M&C International Conference on Mathematics and Computation Supercomputing, Reactor Physics and Nuclear and Biological Application, Avignon, France," .

Forrester A, Bressloff N, Keane A (2006a). "Optimization using surrogate models and partially converged computational fluid dynamics simulations." *Proceedings of the Royal Society A*, **462**(2071), 2177.

Forrester A, Sobester A, Keane A (2007). "Multi-fidelity optimization via surrogate modelling." *Proc. R. Soc. A*, **463**, 3251–3269.

Forrester AIJ, Bressloff NW, Keane AJ (2006b). "Optimization using surrogate models and partially converged computational fluid dynamics simulations." *Proc. R. Soc. A*, **462**, 2177–2204.

Forrester AIJ, Keane AJ (2009). "Recent advances in surrogate-based optimization." *Progress in Aerospace Sciences*, **45**(1-3), 50 – 79. ISSN 0376-0421. doi:DOI:10.1016/j.paerosci.2008.11.001. URL http://www.sciencedirect.com/science/article/B6V3V-4VBM48X-1/2/e11f4604956d4eac00513dca4d8eb0a3.

Forrester AIJ, Keane AJ, Bressloff NW (2006c). "Design and Analysis of "Noisy" Computer Experiments." *AIAA Journal*, **44**.

Ginsbourger D (2009). *Multiples métamodèles pour l'approximation et l'optimisation de fonctions numériques multivariables*. Ph.D. thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne.

Ginsbourger D, Dupuy D, Badea A, Carraro L, Roustant O (2009). "A note on the choice and the estimation of Kriging models for the analysis of deterministic computer models." *Applied Stochastic Models in Business and Industry*, **25**, 115–131.

Ginsbourger D, Le Riche R, Carraro L (2010). *Computational Intelligence in Expensive Optimization Problems*, chapter Kriging is well-suited to parallelize optimization, pp. 131–162. Studies in Evolutionary Learning and Optimization. Springer.

Gramacy RB (2007). "An R Package for Bayesian Nonstationary, Semiparametric Nonlinear Regression and Design by Treed Gaussian Process Models." *Journal of Statistical Software*, **19**(9).

Gramacy RB, Lee HKH (2009). "Adaptive Design and Analysis of Supercomputer Experiments." *Technometrics*, **51**, 130–145.

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

Huang D, Allen T, Notz W, Miller R (2006a). "Sequential Kriging Optimization Using Multiple Fidelity Evaluations." *Sructural and Multidisciplinary Optimization*, **32**, pp. 369–382 (14).

Huang D, Allen T, Notz W, Zheng N (2006b). "Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models." *Journal of Global Optimization*, **34**, 441–466.

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

Jones D, Schonlau M, Welch W (1998). "Efficient Global Optimization of Expensive Black-Box Functions." *Journal of Global optimization*, **13**, 455–492.

Journel A, Huijbregts C (1978). *Mining Geostatistics*. Academic Press, London.

Journel A, Rossi M (1989). "When do we need a trend model in kriging ?" *Mathematical Geology*, **21**(7), 715–739.

Koehler J, Owen A (1996). "Computer Experiments." *Technical report*, Department of Statistics, Stanford University.

Krige D (1951). "A statistical approach to some basic mine valuation problems on the witwatersrand." *J. of the Chem., Metal. and Mining Soc. of South Africa*, **52**(6), 119–139.

Li R, Sudjianto A (2005). "Analysis of Computer Experiments Using Penalized Likelihood in Gaussian Kriging Models." *Technometrics*, **47**(2), 111–120.

Loeppky J, Sacks J, Welch W (2009). "Choosing the Sample Size of a Computer Experiment: A Practical Guide." *Technometrics*, **51**(4), 366–376.

Mardia K, Marshall R (1984). "Maximum likelihood estimation of models for residual covariance in spatial regression." *Biometrika*, **71**, 135–146.

Martin J, Simpson T (2005). "Use of kriging models to approximate deterministic computer models." *AIAA Journal*, **43**(4), 853–863.

Matheron G (1963). "Principles of geostatistics." *Economic Geology*, **58**, 1246–1266.

Matheron G (1969). "Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau." *Technical Report 1*, Ecole Nationale Supérieure des Mines de Paris.

Mebane W, Sekhon J (xxx). "Genetic Algorithm Using Derivatives: The rgenoud package for R." *Journal of Statistical Software*, **xxx**, xxx.

Mockus J (1988). *Bayesian Approach to Global Optimization*. Kluwer academic publishers.

O'Hagan A (2006). "Bayesian analysis of computer code outputs: a tutorial." *Reliability Engineering and System Safety*, **91**(91), 1290–1300.

Omre H (1987). "Bayesian krigingÑMerging observations and qualified guesses in kriging." *Mathematical Geology*, **19**, 25–39.

Park J, Baek J (2001). "Efficient computation of maximum likelihood estimators in a spatial linear model with power exponential covariogram." *Computer Geosciences*, **27**, 1–7.

Picheny V (2009). *Improving accuracy and compensating for uncertainty in surrogate modelling*. Ph.D. thesis, University of Florida and Ecole des Mines de Saint-Etienne.

Queipo N, Verde A, Pintos S, Haftka R (2006). "Assessing the Value of Another Cycle in Surrogate-Based Optimization." In "11th Multidisciplinary Analysis and Optimization Conference," AIAA.

Queipo NV, Haftka R, Shyy W, Goel T, Vaidyanathan R, Tucker P (2005). "Surrogate-based analysis and optimization." *Progress in Aerospace Sciences*, **41**, 1–28.

Rasmussen C, Williams C (2006). *Gaussian Processes for Machine Learning*. the MIT Press, www.GaussianProcess.org/gpml.

Ripley B (1987). *Stochastic Simulation*. Wiley.

Sacks J, Welch W, Mitchell T, Wynn H (1989). "Design and Analysis of Computer Experiments." *Statistical Science*, **4**(4), 409–435.

Saltelli A, Chan K, Scott E (2000). *Sensitivity analysis*. Wiley.

Santner T, Williams B, Notz W (2003). *The Design and Analysis of Computer Experiments*. Springer, New York.

Sasena MJ, Papalambros P, Goovaerts P (2002). "Exploration of Metamodeling Sampling Criteria for Constrained Global Optimization." *Engineering Optimization*, **34**, 263–278. URL http://www.informaworld.com/10.1080/03052150211751.

Schonlau M (1997). *Computer Experiments and Global Optimization*. Ph.D. thesis, University of Waterloo.

Sobol IM (1993). "Sensitivity analysis for non-linear mathematical model." *Math. Modelling Comput. Exp.*, **1**, 407–414.

Stein M (1999). *Interpolation of spatial data, some theory for kriging*. Springer.

Taddy MA, Gramacy RB (2010). "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).

Vazquez E, Bect J (2009). "Convergence properties of the expected improvement algorithm." ArXiv:0712.3744v3.

Welch WJ, Buck RJ, Sacks J, Wynn HP, Mitchell TJ, Morris MD (1992). "Screening, predicting, and computer experiments." *Technometrics*, **34**, 15–25.

Williams BJ, Santner TJ, Notz WI (2000). "Sequential design of computer experiments to minimize integrated response functions." *Statistica Sinica*, **10**, 1133–1152. URL http://www3.stat.sinica.edu.tw/statistica/oldpdf/A10n46.pdf.

Affiliation:

Olivier Roustant
Département 3MI
Ecole des Mines
158, cours Fauriel
42000 St-Etienne, France
E-mail: roustant@emse.fr
URL: http://www.emse.fr/~roustant/

David Ginsbourger
University of Bern
Institute of Mathematical Statistics and Actuarial Science
Alpeneggstrasse 22
CH-3012 Bern, Switzerland
E-mail: david.ginsbourger@stat.unibe.ch

Yves Deville
Alpestat
E-mail: deville.yves@alpestat.com
URL: http://alpestat.com/