Additive Kernels for High-dimensional Gaussian Process Modeling

Nicolas Durrande, David Ginsbourger, Olivier Roustant

Additive Kernels for High-dimensional Gaussian Process Modeling

N. Durrande*‡, D. Ginsbourger†, O. Roustant *

January 12, 2010

Abstract

Gaussian Process (GP) models are often used as mathematical approximations of time expensive numerical simulators. Provided that its kernel is suitably chosen and that enough data is available to obtain a reasonable fit of the simulator, a GP model can be beneficially used for as many tasks as prediction, optimization, or Monte-Carlo-based quantification of uncertainty. However, the former conditions become unrealistic when using classical GPs as the dimension of input increases. One popular alternative is then to turn to Generalized Additive Models (GAMs), relying on the assumption that the simulator's response can approximately be decomposed as a sum of univariate functions. If such an approach has been successfully applied in approximation, it is nevertheless not completely compatible with the GP framework and its versatile applications. The ambition of the present work is to give an insight into the use of GPs for GAMs by integrating additivity within the kernel, and proposing a parsimonious numerical method for data-driven parameter estimation. The first part of this paper deals with the kernels naturally associated to additive processes and the properties of the GP models based on such kernels. The second part is dedicated to a numerical procedure based on relaxation for additive kernel parameter estimation. Finally, the efficiency of the proposed method is illustrated and compared to other approaches on Sobol's g-function in dimension 4, 8 and 12.

keywords: Kriging, Additive Models, Maximum Likelihood Estimation, Relaxed Optimization

*CROCUS - Ecole Nationale Supérieure des Mines de St-Etienne,
29 rue Ponchardier - 42023 St Etienne, France

†CHYN - University of Neuchâtel,
Rue Emile Argand 11 - 2007 Neuchâtel, Switzerland

‡Corresponding author: *durrande@emse.fr*

1 Introduction

The study of numerical simulator often deals with time expensive computer codes. This cost implies that the number of evaluations of the numerical simulator (i.e. the objective function) is limited and thus many methods such as uncertainty propagation, sensitivity analysis, or global optimization are unaffordable. A well known approach is to replace the numerical simulator by a mathematical approximation called metamodel. The metamodel (or response surface or surrogate model) is constructed based on the responses of the simulator for a limited number of inputs called the Design of Experiments (DoE). There is a large number of metamodels types and among the most popular we can cite regression, splines, neural networks... In this paper, we focus on a particular type of metamodel: the Kriging method, more recently referred to as Gaussian Process modeling. Originally presented in spatial statistics [1] as an optimal Linear Unbiased Predictor (LUP) of random processes, Kriging has become a very popular interpolator in machine learning, where its interpretation is usually restricted to the convenient framework of Gaussian Processes (GP). Beyond the LUP —which then elegantly coincides with a conditional expectation—, the latter GP interpretation allows indeed the explicit derivation of conditional probability distributions for the response values at any point or set of points in the input space.

In the case of high dimensional input space $D \subset \mathbb{R}^d$ ($d \in \mathbb{N}\backslash\{0\}$), the classical Kriging method faces two issues: since this method is based on neighborhoods, the first one is to require more and more points in the DoE to cover the domain D. The second one is that the number of anisotropic kernel parameters to be estimated increases with d so that the estimation becomes particularly difficult for high dimensional input spaces [2, 3]. On the other hand, one possibility to get around this issue in state-of-the-art multidimensional statistics is to consider simplified metamodels. A well known type of such metamodels drastically decreasing complexity in high-dimensional approximation is the family of Generalized Additive Models (GAMs). The latter assume indeed that the response can be approximately decomposed as a sum of univariate functions:

$$m(x) = \mu + \sum_{i=1}^{d} m_i(x_i), \tag{1}$$

where $\mu \in \mathbb{R}$ and the m_i's may be linear or non-linear. Since their introduction by Stones in 1985 [4], many methods have been proposed for the estimation of those models. We can cite the method of marginal integration [5] and a very popular method described by Hastie and Tibshirani in [6]: the GAM backfitting algorithm. However, those methods do not consider the probabilistic framework of GP modeling and do not provide additional information such as the prediction variance. Combining the high-dimensional advantages of GAMs with the versatility of GPs is the main goal pursued in the present work.

The first part of this paper focuses on the case of additive Gaussian Processes, their associated kernels and the properties of associated *Additive Kriging*

models. The second part deals with a Relaxed Likelihood Maximization (RLM) procedure for the estimation of kernel parameters for Additive Kriging models. Finally, the proposed algorithm is compared with existing methods on a well known test function: the g-function of Sobol [7]. It is shown within the latter example that Additive Kriging with RLM outperforms both usual Kriging and Additive Kriging with crude MLE, and produce similar performances as Marrel's variant of the Welch algorithm [8]. Furthermore, the obtained approximation results are comparable with GAMs —with fluctuations depending on the dimension— so that the proposed Additive Kriging model is a serious and promising alternative to GAMs in the cases where a built-in probabilistic framework such as a GP model is needed for further applications.

2 Towards Additive Kriging

2.1 Additive random processes

We say that a function $f : D \longrightarrow \mathbb{R}$ is additive when it can be written as $f(x) = \mu + \sum_{i=1}^{d} f_i(x_i)$ where x_i is the i-th component of the d-dimensional input vector x, $\mu \in \mathbb{R}$ is a constant, and the f_i's are arbitrary univariate functions. In this section, we show that interesting additivity properties of square integrable processes can be characterized by some class of kernels. Let us first consider two such independent real-valued first order stationary processes Z_1 and Z_2 defined over the same probability space $(\Omega, \mathcal{F}, P)$ and indexed by $D = \mathbb{R}$, so that their trajectories $Z_i(.;\omega) : x \in D \longrightarrow Z_i(x;\omega)$ are univariate real-valued functions. Let $K_i : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ $(i \in \{1, 2\})$ be their respective covariance kernels and $\mu_1, \mu_2 \in \mathbb{R}$ their means. Then, the process $Z := Z_1 + Z_2$ defined over $(\Omega, \mathcal{F}, P)$ and indexed by $\mathbb{R}^2$, and so that

$$\forall \omega \in \Omega \ \forall x \in \mathbb{R}^2 \ Z(x;\omega) = Z_1(x_1;\omega) + Z_2(x_2;\omega), \tag{2}$$

has mean $\mu = \mu_1 + \mu_2$ and kernel $K(x, y) = K_1(x_1, y_1) + K_2(x_2, y_2)$. Following eq. 2, the latter sum process clearly has additive paths. In this document, we call additive any kernel of the form $K : (x, y) \in (\mathbb{R}^d)^2 \longrightarrow K(x, y) = \sum_{i=1}^{d} K_i(x_i, y_i)$ where the K_i's are positive semidefinite kernels over $\mathbb{R} \times \mathbb{R}$. It is well known [9] that such a combination of low-dimensional semidefinite kernels is also a semidefinite kernel in the direct sum space. Moreover, we show here that the paths of any random process possessing such an additive kernel have interesting properties:

Proposition 1. *Any (square integrable) random process Z_x possessing an additive kernel is additive up to a modification. In essence, it means that there exists a process A_x which paths are all additive, and such that $\forall x \in X$, $\mathbb{P}(Z_x = A_x) = 1$.*

The proof of this property is given in appendix of the paper for $d = 2$ (same pattern for $d \geq 2$ but more cumbersome notations). Note that the class of actual additive processes is not limited to processes with additive kernels. For example,

let us consider Z_1 and Z_2 two correlated Gaussian processes on $(\Omega, \mathcal{F}, P)$ such that the couple (Z_1, Z_2) is Gaussian. Then $Z_1 + Z_2$ is also a Gaussian process with additive paths but its kernel is not additive. However, in the next section, the term additive process will always refer to square integrable processes with additive kernels.

2.2 Invertibility of covariance matrices

In practice, the covariance matrix K of the observations of an additive process Z at a design of experiments $X = (x^{(1)} \ \dots \ x^{(n)})^T$ may not be invertible even if there is no redundant point in X. Indeed, the additivity of Z may introduce linear relationships (that holds almost surely) between the observed values of Z and lead to the non invertibility of K. Figure 1 shows two examples of designs leading to a linear relationship between the observation. For the left panel, the additivity of Z implies that $Z(x^{(4)}) = Z(x^{(2)}) + Z(x^{(3)}) - Z(x^{(1)})$ and thus the fourth column of the covariance matrix is a linear combination of the three other columns: $K(x^{(i)}, x^{(4)}) = K(x^{(i)}, x^{(2)}) + K(x^{(i)}, x^{(3)}) - K(x^{(i)}, x^{(1)})$ and the associated covariance matrix is not invertible.

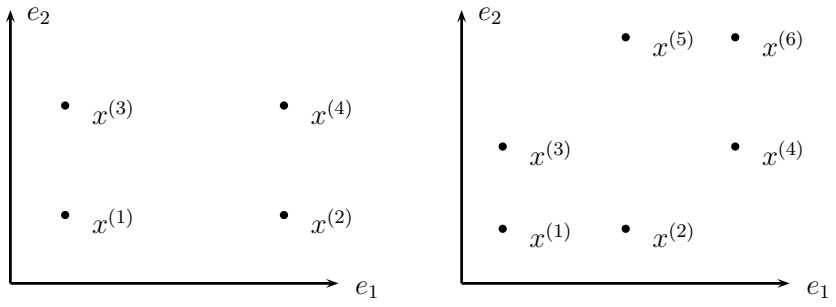

Figure 1: 2-dimensional examples of DoE without replication but which corresponding covariance matrices are systematically non-invertible in the case of random processes with additive kernel.

A first approach is to remove some points in order to avoid any linear combination, which is furthermore in accordance with the aim of parsimonious evaluations for costly simulators. A second approach is to add observation noise in order for K to become invertible.

2.3 Additive Kriging

Let $z : D \mapsto \mathbb{R}$ be the function of interest (a numerical simulator for example), where $D \subset \mathbb{R}^d$. The responses of z at the DoE $\mathcal{X}$ are noted $\mathbf{Z} = (z(x^{(1)}) \ ... \ z(x^{(n)}))^T$. Simple kriging relies on the hypothesis that z is one path of a centered random process Z with kernel K. The expression of the best

predictor (also called kriging mean) and of the prediction variance are :

$$\begin{aligned} m(x) &= k(x)^T \mathrm{K}^{-1}\mathbf{Z} && (3)\\ v(x) &= K(x,x) - k(x)^T \mathrm{K}^{-1} k(x) && (4) \end{aligned}$$

where $k(x) = (K(x,x^{(1)}) \ \dots \ K(x,x^{(n)}))^T$ and K is the covariance matrix of general term $K_{i,j} = K(x^{(i)},x^{(j)})$. Note that these equations correspond respectively with the conditional expectation and variance in the case of a GP with known kernel. In practice however, the structure of k is known (e.g. power-exponential or Matèrn families) but its parameters are unknown. A common way to estimate them is to maximize the likelihood of the kernel parameters given the observations $\mathbf{Z}$ [10, 9]. Let ψ^* denote the parameters maximizing the likelihood.

Equations 3 and 4 are valid for any positive semidefinite kernel. If the covariance is invertible, all usual kriging formulas can be applied for additive processes. Furthermore, the particular structure of their kernel implies some interesting properties:

Proposition 2. *The kriging mean $m(x)$ is an additive function. Indeed,*

$$\begin{aligned} m(x) &= k(x)^T K^{-1}\mathbf{Z} \\ &= (k_1(x_1) + k_2(x_2))^T (K_1 + K_2)^{-1}\mathbf{Z} \\ &= k_1(x_1)^T (K_1 + K_2)^{-1}\mathbf{Z} + k_2(x_2)^T (K_1 + K_2)^{-1}\mathbf{Z} \end{aligned} \quad (5)$$

Another interesting property concerns v: the latter can be null at points that does not belong to the DoE. Let us consider a two dimensional example: the DoE is composed of the 3 points represented on figure 1, $\mathcal{X} = \{x^{(1)} \ x^{(2)} \ x^{(3)}\}$, and we are interested in $v(x^{(4)})$:

$$\begin{aligned} v(x^{(4)}) &= K(x^{(4)},x^{(4)}) - k(x^{(4)})^T \mathrm{K}^{-1} k(x^{(4)}) \\ &= K(x^{(4)},x^{(4)}) - (k(x^{(2)}) + k(x^{(3)}) - k(x^{(1)}))^T \mathrm{K}^{-1} k(x^{(4)}) \\ &= K_1(x_1^{(4)},x_1^{(4)}) + K_2(x_2^{(4)},x_2^{(4)}) - \\ &\qquad (-1 \ \ 1 \ \ 1) \begin{pmatrix} K_1(x_1^{(1)},x_1^{(4)}) + K_2(x_2^{(1)},x_2^{(4)}) \\ K_1(x_1^{(2)},x_1^{(4)}) + K_2(x_2^{(2)},x_2^{(4)}) \\ K_1(x_1^{(4)},x_1^{(4)}) + K_2(x_2^{(3)},x_2^{(4)}) \end{pmatrix} \\ &= K_1(x_1^{(2)},x_1^{(2)}) + K_2(x_2^{(3)},x_2^{(3)}) - K_1(x_1^{(2)},x_1^{(2)}) - K_2(x_2^{(3)},x_2^{(3)}) = 0 \end{aligned}$$

This particularity follows from the fact that the value of the additive process can be known almost surely at the point $x^{(4)}$ based on the observations at $\mathcal{X}$. In the next section, we illustrate the potential of Additive Kriging on an example and propose an algorithm for parameter estimation.

3 Parameter estimation

3.1 Maximum likelihood estimation (MLE)

MLE is a standard way to estimate covariance parameters. This procedure has been covered in detail in the literature [9, 11]. Let Z be a centered additive Process and $\psi_i = \{\sigma_i^2, \theta_i\}$ with $i \in \{1, \ldots, d\}$ the parameters of the univariate kernels. According to the MLE methodology, the best values ψ_i^* for the parameters ψ_i are the values maximizing the likelihood

$$\mathcal{L}(\psi_1, \ldots, \psi_d) := \frac{1}{(2\pi)^{n/2}\det(\mathrm{K}(\psi))^{1/2}} \exp\left(-\frac{1}{2}\mathrm{Y}^T\mathrm{K}(\psi)^{-1}\mathrm{Y}\right) \tag{6}$$

where $\mathrm{K}(\psi) = \mathrm{K}_1(\psi_1) + \cdots + \mathrm{K}_d(\psi_d)$ is the covariance matrix depending on the parameters ψ_i. The latter maximization problem is equivalent to the usually preferred minimization of

$$l(\psi_1, \ldots, \psi_d) := \log(\det(\mathrm{K}(\psi))) + \mathrm{Y}^T\mathrm{K}(\psi)^{-1}\mathrm{Y} \tag{7}$$

Obtaining the optimal parameters ψ_i^* relies on the succesful use of a non-convex global optimization routine. This can be severely hindered for large values of d since the number of kernel parameters increases and so does the search space become high dimensional. One way to cope with this issue is to separate the variables and split the optimization into several low-dimensional subproblems, as we propose in the next subsection.

3.2 The Relaxed Likelihood Maximization algorithm

The aim of the Relaxed Likelihood Maximization (RLM) algorithm is to treat separately the optimization in each direction. In this way, it can be seen as a procedure of optimization via cyclic relaxation [12] with an initial value of the parameters σ_i^2 set to zero. As we will see, the main originality here is to consider a kriging model with an observation noise variance τ^2 that explains the variability in the remaining directions during the optimization.

The first step of the algorithm is to estimate the parameters of the kernel K_1. The simplification of the method is to consider that all the variations of Z in the other directions can be summed up as a white noise. Under this hypothesis, l depends on ψ_1 and τ:

$$l(\psi_1, \tau) = \log(\det(\mathrm{K}_1(\psi_1) + \tau^2 I_d)) + \mathrm{Y}^T(\mathrm{K}_1(\psi_1) + \tau^2 I_d)^{-1}\mathrm{Y} \tag{8}$$

Then, the couple $\{\psi_1^*, \tau^*\}$ that maximizes $L(\psi_1, \tau)$ can be obtained by numerical optimization.
The second step of the algorithm consists in estimating ψ_2, with ψ_1 fixed to ψ_1^*:

$$\begin{aligned} \{\psi_2^*, \tau^*\} =& \underset{\psi_2, \tau}{\mathrm{argmax}}(l(\psi_1^*, \psi_2, \tau)), \text{ with} \\ l(\psi_1^*, \psi_2, \tau) =& \log(\det(\mathrm{K}_1(\psi_1^*) + \mathrm{K}_2(\psi_2) + \tau^2 I_d)) + \\ & \mathrm{Y}^T(\mathrm{K}_1(\psi_1^*) + \mathrm{K}_2(\psi_2) + \tau^2 I_d)^{-1}\mathrm{Y} \end{aligned} \tag{9}$$

This operation can be repeated for all the directions until the estimation of ψ_d. However, even if all the parameters ψ_i have been estimated, it is fruitful to re-estimate them so as the estimation of the parameter ψ_i can benefit of the values ψ_j^* for $j > i$. Thus, the algorithm is composed of a cycle of estimations that treat each direction one after each other:

RLM Algorithm :
1. Initialize the values $\sigma_i^{2(0)} = 0$ for $i \in \{1, \ldots, d\}$
2. For k from 1 to number of iteration do
3. For l from 1 to d do
4. $\{\psi_l^{*(k)}, \tau^{*(k)}\} = \underset{\psi_l, \tau}{\operatorname{argmax}}(\mathcal{L}_c(\psi_1^{*(k)}, \ldots, \psi_{l-1}^{*(k)}, \psi_l, \psi_{l+1}^{*(k-1)}, \ldots, \psi_d^{*(k-1)}, \tau))$
5. End For
6. End For

τ is a parameter tuning the fidelity of the model since for $\tau = 0$ the kriging mean interpolates the data. In practice, this parameter is decreasing at each new estimation. Depending on the observations and on the DoE, it converges either to a constant or to zero (Cf. example and figure 3). When zero is not reached, τ^2 should correspond to the part of variance that cannot be explained by the additive model. Thus, the comparison between τ^2 and the σ_i^* allows us to quantify the degree of additivity of the objective function according to the model.

This procedure of estimation cannot be applied for kernels that are not additive. However, the method developed by Welch for usual kernels in [8] is in the same fashion since it corresponds to a sequential estimation of the parameters. Moreover, one interesting particularity of Welch's algorithm is to choose at each step the best direction in which the parameters should be estimated. The RLM algorithm can easily be adapted in a similar way to improve the quality of the results but the adapted version would be much more time consuming.

4 Application to the g-function of Sobol

In order to illustrate the methodology and to compare it to existing algorithms, an analytical test case is considered. The objective function is the g-function of Sobol defined on $[0, 1]^d$ as

$$g(x_1, \ldots, x_d) = \prod_{k=1}^{d} \frac{|4x_k - 2| + a_k}{1 + a_k} \text{ with } a_k > 0 \tag{10}$$

This function is well known in the literature [7] and it has two main advantages for our study: first, the dimension of the input space is tunable and second, the Sobol sensitivity indices associated to the variables x_i, $i = 1, \ldots, d$ can be

obtained analytically:

$$S_i = \frac{\frac{1}{3(1+a_i)^2}}{\left[\prod_{k=1}^{d} 1 + \frac{1}{3(1+a_k)^2}\right] - 1} \tag{11}$$

Obviously, this function is not additive but depending on the coefficients a_k, it can be very closed to an additive function. For example the sum of the first order Sobol indices is 0.99 for $a_k = 5$ and $d = 4$. As a rule, the g-function is all the more additive as the a_k are large

This section presents 3 tests in dimension 4, 8 and 12. In order to compare our methodology with the results of [13], we choose $a_k = k$. The DoE are LH maximin designs [11] composed of $10 \times d$ points. To asses the quality of the obtained models, the predictivity coefficient Q_2 is computed on a test sample of $n = 1000$ points uniformly distributed on $[0, 1]^d$. Its expression is:

$$Q_2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

where y is the vector of observation, $\hat{y}$ is the vector of predicted values and $\bar{y}$ is the mean of y.

4.1 A 4-dimensional example

In 4 dimension with the coefficients $a_k = k$, the sum of the Sobol indices of the main effects is equal to 0.95 so the g-function is almost additive. We run on this example 5 iterations of the RLM algorithm. The univariate functions obtained are presented on figure 2. On this figure, the represented values are centered to ensure that the observations and the univariate functions are comparable. This example highlights one particular advantage of additivity: since the effect of any variable can be isolated and represented, the model is easily interpretable.

The evolution of the estimated observation noise variance τ^2 is represented on figure 3. On this figure, it appears that the convergence of the algorithm is reached at the iteration 4 and the final τ^2 is 0.01. In this example, the quality of the constructed model is high since coefficient of prediction Q_2 on the test sample is equal to 0.91. Nevertheless, this depends on the DoE and on the optimization performances, so that it is likely to fluctuate.

4.2 Comparison with other algorithms

In this section, the proposed RLM algorithm is compared with other methods: additive kriging model with a "crude" MLE, kriging with usual kernel, the GAM algorithm, and a method of GP modeling developed by A. Marrel in [13]. The results for usual kriging are obtained with the DiceKriging R package [14], and the GAM implementation comes with the R package available on the CRAN website. Eventually, the results of A. Marrel presented in [13] are cited in order

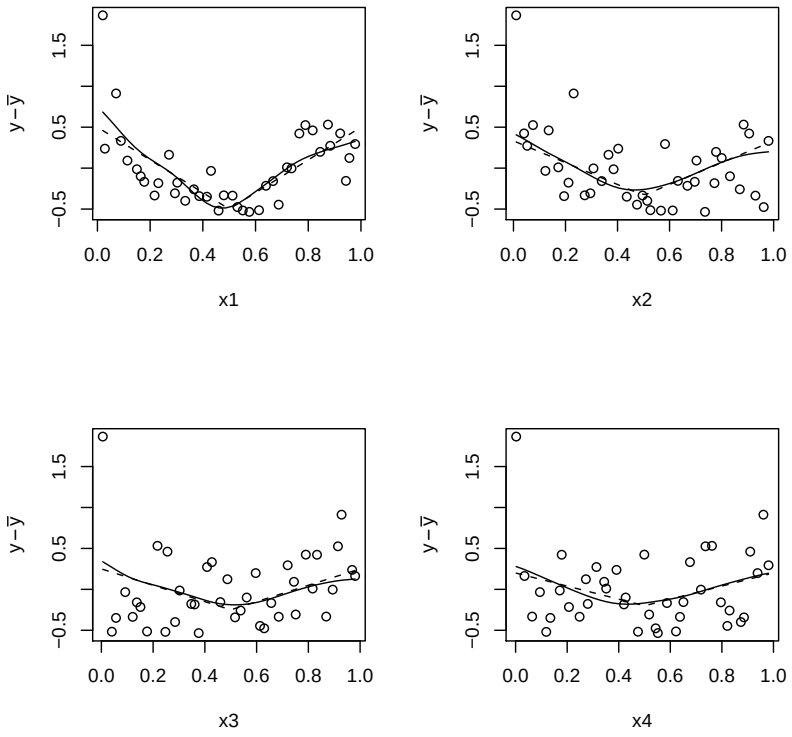

Figure 2: 1-dimensional projections of the observations (bullets) on the g-function example with $d = 4$. The univariate models (solid lines) obtained after 5 iterations of RLM are very closed to the analytical main effects (dashed lines).

to compare our methodology to a specific algorithm for high dimensional kriging model selection.

As previously, the test function is the g-function and the coefficients are set to $a_k = k$. With this values, the sum of the first order Sobol indices varies from 0.95 for $d = 4$ to 0.93 for $d = 12$. For the methods RLM, MLE and DiceKriging, the procedure used for the likelihood maximizations is the BFGS algorithm from the *optim* R package. The obtained Q_2's are gathered in table 1.

This example deserves many remarks. First, the results are based on the construction of a single model and they are sensitive to many parameters such as the DoE and the robustness of the optimization. Experience shows that the results obtained by RLM are less robust than the results from DiceKriging and

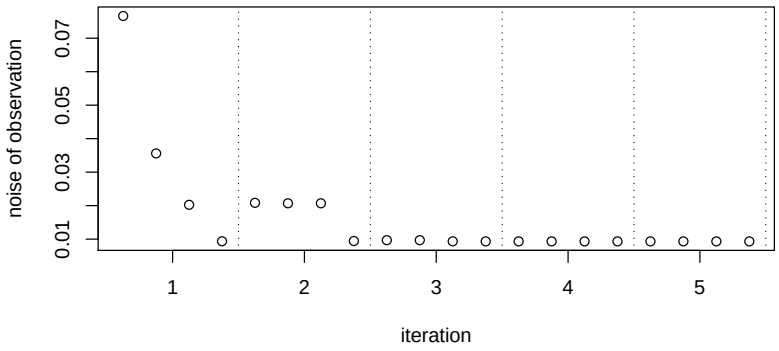

Figure 3: Evolution of the observation noise on the 4-dimensional example

Algorithm	kernel	4	8	12
RLM	Additive Matern 3/2	**0.91**	0.79	0.72
RLM	Additive Matern 5/2	0.86	0.75	**0.85**
MLE	Additive Matern 3/2	0.86	**0.82**	0.63
DiceKriging	Matern 3/2	0.79	0.66	0.67
Marrel	power-exponential	0.86	**0.85**	**0.84**
GAM	(smoothing splines)	**0.89**	0.81	0.74

Table 1: Q_2 predictivity coefficients at a 1000-points test sample for the various methodologies for $d \in \{4, 8, 12\}$.

GAM but more than those from MLE with additive kernel. Our current implementation of RLM and MLE may be at stake and future work will focus on its improvement. A second remark is that the RLM algorithm with additive Matern 3/2 kernel gives good results in low dimension whereas the additive Matern 5/2 kernel performs better in higher dimension. This fact is not surprising since the g-function with $a_k = k$ is all the more smooth in the k-th direction as k is large. But this choice for the a_k implies also that the influence of the variable x_k decreases with k. This assumption of significant and non-significant variables is advantageously used by Marrel's algorithm, and probably explains its robustness regarding dimensionality.

5 Concluding remarks

The proposed methodology seems to be a good challenger for high-dimensional kriging modeling. On the presented example, it takes advantage of the impor-

tant additive component of the g-function, so that the best metamodels are almost systematically found among those based on additive kernels. Furthermore, RLM outperforms crude MLE in most cases. In all, the proposed models take advantage of additivity, while benefiting from interesting features inherited from both the kriging and dimension reduction frameworks. The latters include the versatility of a GP model, but also the flexibility of choosing the univariate kernels, like for the predictors in GAM.

One advantage of the RLM algorithm compared to the backfitting proposed in the *gam* package is that it estimates the covariance structure in each direction. Since the smoothing parameters are estimated at the beginning of the GAM algorithm and do not vary afterward, the RLM is most likely to give interesting results by adapting the kernel at each step. The proposed methodology hence seems very promising, even if many points still require a particular improvement. Among them, we can cite the construction of a stopping criteria for the cycle of estimation, and more ambitiously some proof of convergence for the RLM. In other respects, experimental works to be conducted in a near future include improving the robustness of likelihood optimizers used in RLM, and test the forthcoming enhanced RLM in higher dimensions.

References

[1] N. Cressie, Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics, 1993.

[2] K.-T. Fang, R. Li, A. Sudjianto, Design and modeling for computer experiments, Chapman & Hall, 2006.

[3] A. OHagan, Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering and System Safety 91.

[4] C. Stone, Additive regression and other nonparametric models, The annals of Statistics 13 (2) (1985) 689–705.

[5] W. Newey, Kernel estimation of partial means and a general variance estimator, Econometric Theory (1994) 233–253.

[6] T. Hastie, R. Tibshirani, Generalized Additive Models, Monographs on Statistics and Applied Probability, 1990.

[7] A. Saltelli, K. Chan, E. Scott, Sensitivity analysis, Wiley Series in Probability and Statistics, 2000.

[8] W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, M. D. Morris, Screening, predicting, and computer experiments, Technometrics 34 (1992) 15–25.

[9] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[10] D. Ginsbourger, D. Dupuy, A. Badea, O. Roustant, L. Carraro, A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments, Applied Stochastic Models for Business and Industry 25 (2009) 115 –131.

[11] T. J. Santner, B. Williams, W. Notz, The Design and Analysis of Computer Experiments, Springer-Verlag, 2003.

[12] M. Minoux, Mathematical Programming: Theory and Algorithm, John Wiley & Sons, 1986.

[13] A. Marrel, B. Iooss, F. Van Dorpe, E. Volkova, An efficient methodology for modeling complex computer codes with gaussian processes, Computational Statistics and Data Analysis, 52 52 (2008) 4731–4744.

[14] O. Roustant, D. Ginsbourger, Y. Deville, The DiceKriging package: kriging-based metamodeling and optimization for computer experiments, in: Book of abstract of the R User Conference. Package available at www.dice-consortium.fr, 2009.

Appendix: Proof of proposition 1 for $d = 2$

Let Z be a random process indexed by $\mathbb{R}^2$, with kernel $K(x, y) = K_1(x_1, y_1) + K_2(x_2, y_2)$, and Z_T the random process defined by $Z_T(x_1, x_2) = Z(x_1, 0) + Z(0, x_2) - Z(0, 0)$. By construction, the paths of Z_T are additive functions. In order to show the additivity of the paths of Z, we will show that $\forall x \in \mathbb{R}^2$, $\mathrm{P}(Z(x) = Z_T(x)) = 1$. For a sake of simplicity, the three terms of $\mathrm{var}[Z(x) - Z_T(x)] = \mathrm{var}[Z(x)] + \mathrm{var}[Z_T(x)] + \mathrm{cov}[Z(x), Z_T(x)]$ are studied separately:

$$\mathrm{var}[Z(x)] = K(x, x)$$

$$\begin{aligned}
\mathrm{var}[Z_T(x)] &= \mathrm{var}[Z(x_1, 0) + Z(0, x_2) - Z(0, 0)] \\
&= \mathrm{var}[Z(x_1, 0)] + \mathrm{var}[Z(0, x_2)] + \mathrm{var}[Z(0, 0)] + 2\mathrm{cov}[Z(x_1, 0), Z(0, x_2)] \\
&\quad - 2\mathrm{cov}[Z(x_1, 0), Z(0, 0)] - 2\mathrm{cov}[Z(0, x_2), Z(0, 0)] \\
&= K_1(x_1, x_1) + K_2(0, 0) + K_1(0, 0) + K_2(x_2, x_2) + K_1(0, 0) + K_2(0, 0) \\
&\quad + 2\left(K_1(x_1, 0) + K_2(0, x_2) - K_1(x_1, 0) - K_2(0, 0) - K_1(0, 0) - K_2(x_2, 0)\right) \\
&= K_1(x_1, x_1) + K_2(x_2, x_2) = K(x, x)
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}[Z(x), Z_T(x)] &= \mathrm{cov}[Z(x_1, x_2), Z(x_1, 0) + Z(0, x_2) - Z(0, 0)] \\
&= K_1(x_1, x_1) + K_2(x_2, 0) + K_1(x_1, 0) + K_2(x_2, x_2) \\
&\quad - K_1(x_1, 0) - K_2(x_2, 0) \\
&= K_1(x_1, x_1) + K_2(x_2, x_2) = K(x, x)
\end{aligned}$$

Those three equations implies that $\mathrm{var}[Z(x) - Z_T(x)] = 0$, $\forall x \in \mathbb{R}^2$. Thus, $\mathrm{P}(Z(x) = Z_T(x)) = 1$ and there exist a modification of x with additive paths.