

Data-driven Kriging models based on FANOVA-decomposition

Thomas Muehlenstaedt¹, Olivier Roustant², Laurent Carraro³ and Sonja Kuhnt¹

¹ Faculty of Statistics
TU Dortmund University
Dortmund, Germany
{muehlens, kuhnt}@statistik.tu-dortmund.de

² LSTI / CROCUS team
Ecole des Mines de Saint-Etienne
Saint-Etienne, France
roustant@emse.fr

³ LAMUSE
Telecom Saint-Etienne
Saint-Etienne, France
laurent.carraro@telecom-st-etienne.fr

November 2010

Abstract

The situation of time consuming computer experiments is considered, where the output is deterministic and the data generating function is of high complexity. In such situations the underlying functions often are non additive but at the same time, not all interactions are active. Hence neither a model considering all interactions as well as an additive model is adequate. As a solution a modified Kriging model is proposed, which reflects the interaction structure inherent to the data generating mechanism. This is achieved by exploring the interaction structure of the output based on FANOVA methods. For illustrating the interaction structure, a graph is developed which summaries the structure of the output generating function in additive parts. Finally, modified covariance kernels are defined, which allow for a more precise modeling of simulation output.

Keywords. Sensitivity Analysis, Computer Experiment, Functional Decomposition, Graph, Kriging

1 Introduction

For many phenomena there exist time consuming simulation models which are capable of predicting the output of real world experiments very precisely and are thus used as a replacement for real experiments. These simulation models are often deterministic such that repetitions and randomization are not appropriate for designing simulation experiments. Working with simulation models one often has to deal with the constraint that due to computation time only very limited runs of the simulation are available. Hence conducting simulation experiments require careful planing which simulation runs should be realized. As normally not all combinations of input variables of interest are available, fast models for predicting the simulation output at untried design points are desirable. The standard model in this situation is Kriging, see e.g. [15] and [7], which is capable of modeling highly complex data and also can be used as interpolation method.

As a motivation for constructing a Kriging prediction model consider the so called Ishigami function, which is a popular function for illustrating sensitivity analysis, see e.g. [13]:

$$f(x) = \sin(x_1) + A \sin^2(x_2) + B x_3^4 \sin(x_1) \quad (1)$$

with $x \in [-\pi, \pi]^3$ and $A = 7, B = 0.1$. The Ishigami function is often chosen as it has a high complexity including relevant interaction terms. Assuming that the function is unknown and we have observed 100 runs of the function, the aim is to construct

a prediction model for the unknown function. One popular choice is Kriging, also called Gaussian process modeling, which assumes that the observations are drawn from a gaussian random field:

$$Y(x) = \mu + Z(x), \quad (2)$$

with $Z(x)$ being a Gaussian process. The prediction function is than the conditional mean given the observation. A key part about Kriging is the covariance of $Z(x)$. A standard approach is to assume a stationary, anisotropic structure:

$$\text{cov}(Z(x^{(1)}), Z(x^{(2)})) = \sigma^2 \prod_{k=1}^d g_k(x_k^{(1)} - x_k^{(2)}; \theta_k), \quad (3)$$

where θ is a correlation parameter (-vector). The function $g_k(h; \theta_k)$ is a one-dimensional correlation function only depending on the k -th input variable. The parameters μ, σ, θ have to be estimated from the sample. Implicitly this covariance structure assumes that all possible interactions are (at least at a very small scale) active. However, in case of the Ishigami function, we observe, that there is a special structure of the function. It is neither a pure additive model nor it is a function, where all interactions are active. Hence modifying the covariance structure up to this special structure might yield a better fit to the data. Therefore we assume the following covariance structure:

$$\text{cov}(Z(x^{(1)}), Z(x^{(2)})) = \sigma_1^2 \prod_{k \in \{1,3\}} g_k(x_k^{(1)} - x_k^{(2)}; \theta_k) + \sigma_2^2 g_2(x_2^{(1)} - x_2^{(2)}; \theta_2). \quad (4)$$

Assuming this covariance structure, again all parameters can be estimated and predictions can be made. Fitting both models

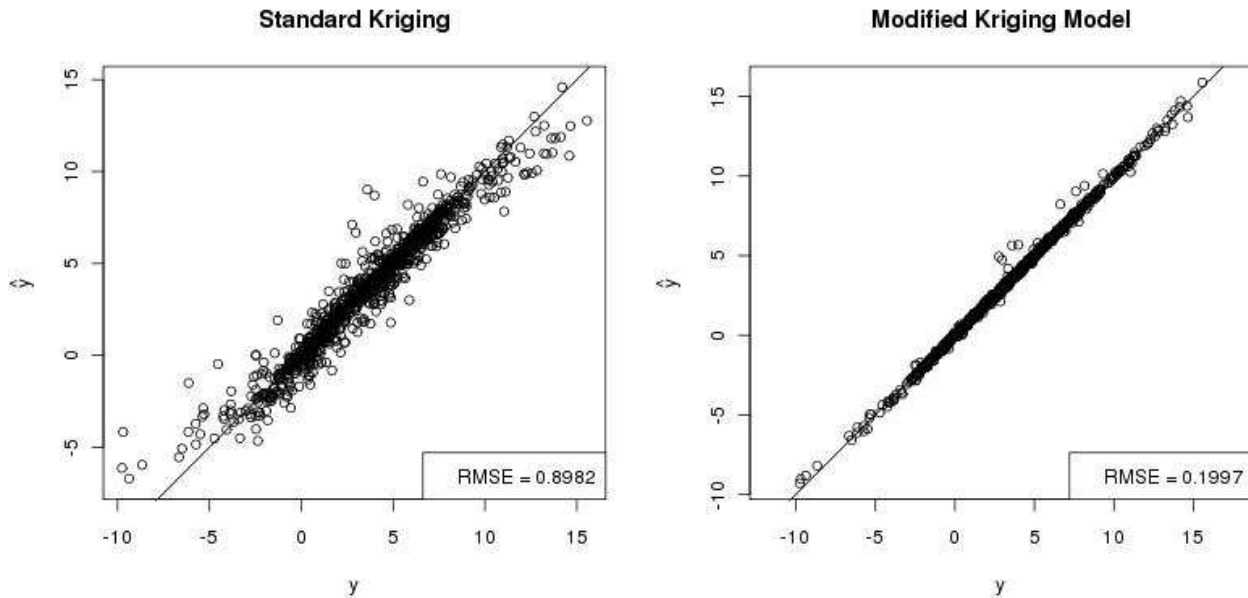


Figure 1: Prediction plots for the Ishigami function. On the left hand plot the result for a standard Kriging model is given, on the right hand side the result for Kriging model with modified covariance function.

(with $\rho_\theta(\cdot)$ being a Matern 5/2 correlation function) to the same 100 observations, predictions are made for 1000 additional observations. The predictions can be compared with the true observations. In figure 1 the prediction power of both models is compared, with the result that the modified model greatly improves the fit. Hence it looks attractive to try to fit models to that data, which are more sensitive to the data than for example standard Kriging models. Two major tools for doing so are used: The functional decomposition introduced for example by [16] and [6] and considered in a stochastic framework for example by [9] and mathematical graphs.

The article is structured as follows: First theory about functional decompositions, graphs and Kriging is revised. Then estimation issues are addressed in section 3. The modified Kriging models are applied in section 4 and a discussion about advantages, disadvantages and applications to areas other than prediction is done in section 5. An outline is concluding the article.

2 Theory

The aim of this section is to introduce new Kriging models, built from a relevant combination of kernels (2.3). Such combinations are based on the cliques of “FANOVA-graphs”, introduced in 2.2, in which vertices represent variables and edges correspond to the presence of (any order) interactions. The first subsection is recalling the main concept of the FANOVA decomposition.

2.1 Functional ANOVA

Consider a continuous function $f : \Delta \rightarrow \mathbb{R}$, $f \in L_2(\Delta, \mathbb{R})$ with $\Delta = \Delta_1 \times \dots \times \Delta_d$. Let X be a random vector over the domain Δ with integration measure $d\nu$. We assume that X_1, \dots, X_d are independent, i.e. that $d\nu = d\nu_1 \dots d\nu_d$. Consider a function f such that the random variable $f(X)$ is square integrable. Then we have the so-called Functional ANOVA (FANOVA) decomposition (see [6] or [16]):

$$f(X) = \mu_0 + \sum_{i=1}^d \mu_i(X_i) + \sum_{j<k} \mu_{jk}(X_j, X_k) + \sum_{j<k<l} \mu_{jkl}(X_j, X_k, X_l) + \dots + \mu_{12\dots d}(X_1, X_2, \dots, X_d). \quad (5)$$

where each term is centered and orthogonal:

$$E(\mu_J(X_J)) = 0 \quad (6)$$

$$\forall J' \neq J : E(\mu_J(X_J)\mu_{J'}(X_{J'})) = 0 \quad (7)$$

In the equations above, we have used the usual index set notation. For instance with $J = \{1, 2\}$, X_J means (X_1, X_2) , and μ_J means $\mu_{1,2}$.

The functions $\mu_1(x_1), \dots, \mu_d(x_d)$ can be interpreted as main effects, and the terms $\mu_{j,k}(x_j, x_k)$, $j < k$ as twofold interactions. This functional decomposition can be uniquely obtained by recursive integration:

$$\mu_0 = E(f(X)), \quad (8)$$

$$\mu_k(x_k) = E(f(X)|X_k = x_k) - \mu_0, \quad (9)$$

$$\mu_{jk}(x_j, x_k) = E(f(X)|X_j = x_j, X_k = x_k) - \mu_j(x_j) - \mu_k(x_k) - \mu_0 \quad (10)$$

and more generally:

$$\mu_J(x_J) = E(f(X)|X_J = x_J) - \sum_{J' \subsetneq J} \mu_{J'}(x_{J'}). \quad (11)$$

Based on the FANOVA decomposition, sensitivity indices can be defined and are interpreted in an analogue manner as for standard ANOVA [16]. The overall variance of the function is given by:

$$D = \text{var}(f(X)) = E(f(X)^2) - \mu_0^2. \quad (12)$$

Now for each term μ_J a similar variance can be obtained:

$$D_J = \text{var}(\mu_J(X_J)). \quad (13)$$

As for standard Least squares ANOVA a variance decomposition holds:

$$D = \sum_{k=1}^d \sum_{|J|=k} D_J. \quad (14)$$

Hence it makes sense to norm the sensitivity indices to

$$S_J = \frac{D_J}{D}. \quad (15)$$

The sensitivity indices S_J are an attractive tool for investigating a function f as they do not require limiting assumptions. Although they are difficult to calculate analytically, they can be calculated numerically [13].

2.2 FANOVA graphs

Graphs are used in a wide range of mathematical fields and are described for example in [3]. In statistics graphs are used in different contexts, e.g. for variable selection [1] and for modeling dependence structure of random vectors [5]. A graph $G = (V, E)$ is a finite set of vertices V and a set of edges combining the vertices in V . The elements in V are indexed by $V = \{1, \dots, d\}$ and E is a set of pairs of vertices from V , which specify the edges of the graph. A concept regarding graphs which is used in the following is the clique. A clique C is a subgraph of G which is complete and which loses the completeness if another vertex is added to C . This can be illustrated using figure 2. For the left hand example there are two cliques $C_1 = \{1, 2, 3\}$, $C_2 = \{4, 5, 6\}$, whereas for the right hand example there are three cliques: $C_1 = \{1, 2, 3\}$, $C_2 = \{4, 5, 6\}$, $C_3 = \{3, 4\}$. For example, for both graphs, the subgraph defined by the set of vertices $\{5, 6\}$ is not a clique, since it is possible to obtain a larger complete subgraph by adding vertex 4. As we only consider undirected graphs, we just state (j, k) , $j < k$ in the set of edges instead of (j, k) and (k, j) .

The intention for the following analysis is, that in general there are $2^d - 1$ terms in the functional decomposition. Even for medium values of d , e.g. $d = 5$, this is a huge amount of data. Therefore most articles on Sobol' indices only consider main

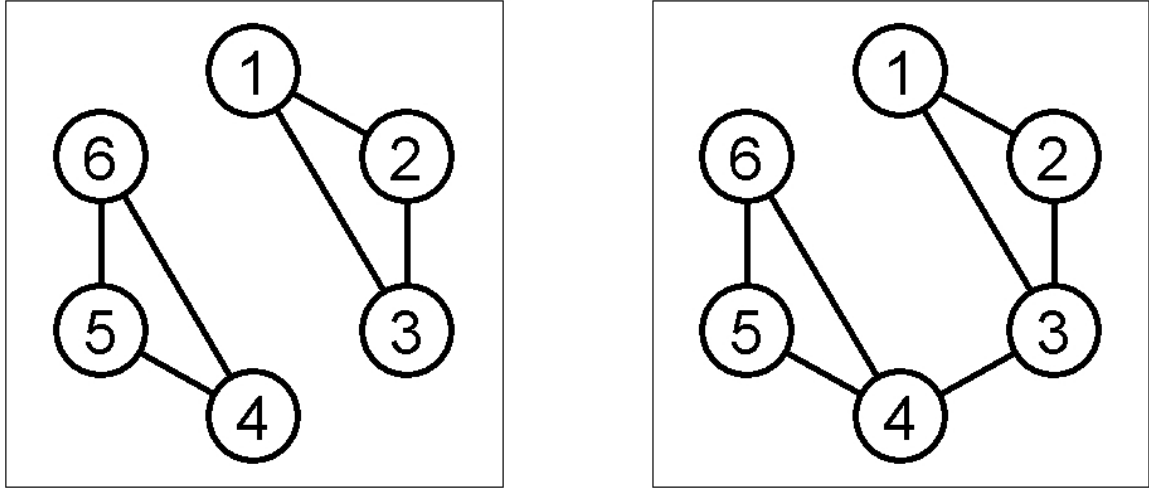


Figure 2: Example of two graphs with $V = \{1, \dots, 6\}$.

effects. Here a methodology is suggested, which reduces data but still gives good insight in the interaction structure of the function f by using graphs.

The set of vertices V is set to be $\{1, \dots, d\}$, such that each vertex represents one input variables. The basic idea is that two vertices / input variables j, k are connected in the graph G if there is any term index set J which includes j, k with $\mu_J(x_J) \neq 0$. More precisely:

Definition 1. (FANOVA graph)

$$(j, k) \in E \Leftrightarrow \exists \mu_J(x_J) \neq 0 \text{ with } \{j, k\} \in J, j \neq k \quad (16)$$

and vice versa

$$(j, k) \notin E \Leftrightarrow \forall J \text{ with } \{j, k\} \in J \text{ it holds that } \mu_J(x_J) \equiv 0. \quad (17)$$

Definition 1 is equivalent to stating that an edge (j, k) is not part of the graph iff for all Sobol indices S_J with $(j, k) \subset J$ it holds: $S_J = 0$. Hence the graph illustrates the parts of the function f which are purely additive. If a graph with cliques C_1, \dots, C_L holds for function f then the function f is additive in the cliques.

Using the information of the graph the functional decomposition becomes

$$f(x) = \mu_0 + \sum_{l=1}^L \psi_{C_l}(x_{C_l}), \quad (18)$$

with

$$\psi_{C_l}(x_{C_l}) := \sum_{I \subset C_l} \mu_I(x_I). \quad (19)$$

This can be simplified when inserting directly the conditional expectations:

$$f(x) = \sum_{l=1}^L E(f(X) | X_{C_l} = x_{C_l}). \quad (20)$$

Remark and assumption. In general, the equalities 18 and 20 hold only ν -almost surely. Nevertheless, it is well known in measure theory that for continuous functions, almost sure equalities are true equalities in the support of ν . Thus, 18 and 20 will hold for all x in Δ if f is a continuous function and the support of ν is equal to Δ (since the continuity of f implies the continuity of the ψ_{C_l} 's). Obviously, such an assumption is not restrictive for most practical applications, and we will consider it from now on.

Note that under this hypothesis, the FANOVA graph will not depend on ν since 5 is also a true equality on the support of ν , and thus on Δ (second part of our assumption). In particular the fact that one μ_J is identically zero then depends entirely on the form of the function f and not on the integration measure.

2.3 Building Kriging models from FANOVA graphs

The information contained in the graph can be used to construct situation specific covariance functions for a Kriging model. In the context of computer experiments, Kriging is a standard tool to predict expensive to evaluate functions at untried locations (see e.g. [15], [7]). For Kriging, the assumed model is:

$$Y(\mathbf{x}) = \sum_{k=1}^p \beta_k f_k(\mathbf{x}) + Z(\mathbf{x}) \quad (21)$$

where $\sum_{k=1}^p \beta_k f_k(\mathbf{x})$ represents the trend value at location \mathbf{x} and $Z(\cdot)$ is a centered Gaussian process with covariance function, or kernel K . We assume that $Z(\cdot)$ is stationary, which implies that K depends only on the difference between two locations, and thus can be written as $K(\mathbf{x}^1, \mathbf{x}^2) = k(\mathbf{x}^1 - \mathbf{x}^2)$ with $k(\cdot) = \sigma^2 R(\cdot; \theta)$, where σ^2 is the process variance, R the correlation function and θ a vector of parameters.

A reason for the success of Kriging is that it interpolates the data, which is desired for deterministic simulators, and, due to its probabilistic nature, also gives a measure of ignorance at unknown points. In this paper, we will focus on the interpolator, given by:

$$\hat{Y}(\mathbf{x}) = \sum_{k=1}^p \beta_k f_k(\mathbf{x}) + \mathbf{k}(\mathbf{x})' \mathbf{K}^{-1} (\mathbf{y} - \mathbf{F}\beta) \quad (22)$$

where \mathbf{x} is a new point at which to predict, \mathbf{K} is the covariance matrix at data, $\mathbf{k}(\mathbf{x})$ the covariance vector between data and \mathbf{x} , and \mathbf{F} the design matrix containing the trend values at data points. In practice, the parameters β_k , σ^2 and θ are estimated and plugged in equation (22).

The function f considered above is understood as a realization of the Gaussian process $Y(\cdot)$. As we see in (21), the departure from the trend relies on the kernel k . Therefore, it is of primary importance to specify it properly. In computer experiments, kernels are often obtained as tensor products of 1-dimensional kernels:

$$k(\mathbf{h}) = \sigma^2 \prod_{k=1}^d g_k(h_k; \theta_k) \quad (23)$$

Famous 1-dimensional covariance functions are the Gaussian one, $g(h; \theta) = \exp\left(-\frac{h^2}{2\theta^2}\right)$ and the Matérn 5/2 one : $g(h; \theta) = \left(1 + \frac{\sqrt{5}|h|}{\theta} + \frac{5h^2}{3\theta^2}\right) \exp\left(-\frac{\sqrt{5}|h|}{\theta}\right)$. Their differentiability properties are directly linked to the smoothness of the sample functions of Z (in mean square sense): existence of derivatives at any order with the Gaussian covariance, existence of second order derivatives with Matérn 5/2. In practice, the Matérn 5/2 choice may correspond to more realistic assumptions on f , and is sometimes recommended [17].

Now in our framework, we can specify kernels to take advantage of the additional knowledge given by the FANOVA decomposition. Assume that a graph $G_{f,\nu}$ is given, with cliques C_1, \dots, C_L . Then, Equation (18) leads to consider the following model for $Z(\cdot)$:

$$Z(\mathbf{x}) = \sum_{l=1}^L Z_{C_l}(\mathbf{x}_{C_l}) \quad (24)$$

where the $Z_{C_l}(\cdot)$ are independent centered Gaussian stationary processes.

Hence, equation 24 implies an additive clique decomposition for the kernel of Z ,

$$\text{cov}(Z(\mathbf{x}^1), Z(\mathbf{x}^2)) = \sum_{l=1}^L \text{cov}(Z_{C_l}(\mathbf{x}_{C_l}^1), Z_{C_l}(\mathbf{x}_{C_l}^2)) \quad (25)$$

which leads to consider kernels of the form

$$k(\mathbf{h}) = \sum_{l=1}^L k_{C_l}(\mathbf{h}_{C_l}) \quad (26)$$

where each k_{C_l} is a kernel defined on the subset of input variables given by the clique C_l . Even if it can make sense in some context, we will assume a common type for all these kernels (for instance Matérn 5/2).

3 Estimation methodology

We now give some insights how to achieve the data-driven construction of Kriging models based on FANOVA graphs in practice (see 2.2 and 2.3). The two different estimation problems (estimating the graph and the new Kriging models) are first addressed separately, for a general function (Sections 3.1 and 3.2). Finally the global estimation procedure is explained in the case of an expensive-to-evaluate function (Section 3.3).

3.1 Graph estimation

As it is seldom the case that the Sobol' indices can be calculated analytically they have to be calculated numerically. However, based on Monte Carlo methods it is very demanding to calculate all Sobol' indices and is therefore not applicable for estimating the graph of a function f . Hence another method is applied here that basically checks additivity of a two dimensional projection of the function.

Recall from 2.2 that f is assumed to be a continuous function on a domain Δ equal to the support of the integration measure ν . This is to guarantee that the equalities below will hold everywhere on Δ and not only almost surely. Now, for any $x_{-(j,k)}$, we consider the function of the two variables x_j and x_k ,

$$f_{x_{-(j,k)}} : (x_j, x_k) \rightarrow f(x). \quad (27)$$

It's FANOVA decomposition can be written as (with analogous notation as in 5):

$$f_{x_{-(j,k)}}(x_j, x_k) = \mu_{0;x_{-(j,k)}} + \mu_{j;x_{-(j,k)}}(x_j) + \mu_{k;x_{-(j,k)}}(x_k) + \mu_{jk;x_{-(j,k)}}(x_j, x_k). \quad (28)$$

This decomposition of course depends on the value $x_{-(j,k)}$, but nevertheless a qualitative statement can be made about the structure of the functional decomposition:

Proposition 1

For the functional decomposition of equation (28) it holds that the edge (j, k) is inactive if and only if for all $x_{-(j,k)}$, $f_{x_{-(j,k)}}$ is additive, i.e. $\mu_{jk;x_{-(j,k)}}(x_j, x_k) \equiv 0$.

Proof. Without loss of generality, let us specify $j = 1$ and $k = 2$. Remark that equation (28) can be rewritten under the form:

$$f_{x_3, \dots, x_d}(x_1, x_2) = f_0(x_3, \dots, x_d) + f_1(x_1, x_3, \dots, x_d) + f_2(x_2, x_3, \dots, x_d) + f_{1,2}(x_1, x_2, x_3, \dots, x_d)$$

where, by definition, $f_{x_3, \dots, x_d}(x_1, x_2) = f(x_1, \dots, x_d)$. Thus f_{x_3, \dots, x_d} is additive for all x_3, \dots, x_d if and only if there is no term depending both on x_1 and x_2 in f , which is equivalent to say that $(1, 2)$ is inactive in the graph \square

Hence it makes sense to use the interaction term of the two-dimensional projections as an indicator whether or not to include an edge into the graph. As a measure of importance, the un-normalized Sobol index of the interaction term of the two-dimensional projection is used:

$$D_{jk}(x_{-(j,k)}) = \text{var}(\mu_{jk;x_{-(j,k)}}(X_j, X_k)). \quad (29)$$

This Sobol index is a function of $x_{-(j,k)}$. In order to have a measure for assessing if an edge is active, integrate $D_{jk}(X_{-(j,k)})$ w.r.t $X_{-(j,k)}$:

Definition 2

$$\mathfrak{D}_{j,k} := E(D_{jk}(X_{-(j,k)})). \quad (30)$$

Obviously $\mathfrak{D}_{j,k} \geq 0$. For the index $\mathfrak{D}_{j,k}$ the following proposition holds:

Proposition 2

Given function f with corresponding graph $G_{f,\nu}$, it holds that

$$\mathfrak{D}_{j,k} > 0 \Leftrightarrow (j, k) \in E. \quad (31)$$

Proof. This result is a direct consequence of Proposition 1 \square

Example 1 For the Ishigami function with independent integration measure, $\mu_{1,2;x_3} = \mu_{2,3;x_1} \equiv 0$, and $\mu_{1,3;x_2}(x_1, x_3) = B(x_3^4 - E(X_3^4))(\sin(x_1) - E(\sin(X_1)))$. Thus, $\mathfrak{D}_{1,2} = \mathfrak{D}_{2,3} = 0$, while $\mathfrak{D}_{1,3} = B^2 \text{var}(X_3^4) \text{var}(\sin(X_1))$. Remark that these indices are the same as the usual (un-normalized) 2nd order Sobol indices. This is because the Ishigami function involves only 2nd order (non linear) interactions.

Example 2 Now consider the function $g(x_1, \dots, x_d) = x_1 x_2 x_3$, again with independent integration measure and with $d \geq 3$. It is easy to see that $\mathfrak{D}_{j,k} = 0$ if $j > 3$ or $k > 3$. Consider $j = 1$ and $k = 2$. We have $\mu_{1,2;x_{-(1,2)}}(x_1, x_2) = (x_1 - E(X_1))(x_2 - E(X_2))x_3$. Thus, $D_{1,2;x_{-(1,2)}} = \text{var}(X_1) \text{var}(X_2) x_3^2$, which depends on x_3 . Finally, $\mathfrak{D}_{1,2} = \text{var}(X_1) \text{var}(X_2) E(X_3^2)$. $\mathfrak{D}_{1,3}$ and $\mathfrak{D}_{2,3}$ are computed in the same way. These results are different from the usual 2nd order Sobol indices $D_{1,2}, D_{1,3}$ and $D_{2,3}$, that are all null.

The idea of fixing variables in the FANOVA is close to the ‘‘cut-HDMR expansion’’ (see e.g. [13], chapter 9). Cut-HDMR is equivalent to a FANOVA decomposition where the integration measure is a product of one-dimensional Dirac-measures. In the above methodology, equation (3.1) is actually the FANOVA decomposition obtained with integration measure $d\nu_j d\nu_k \prod_{s \notin \{j,k\}} d\delta_{x_s}$. In a sense, it is thus intermediate between the usual FANOVA decomposition and cut-HDMR. With this approach, there are only $\binom{d}{2}$ indices $\mathfrak{D}_{j,k}$ in contrast to $2^d - 1$ Sobol indices. If it would be possible to calculate $\mathfrak{D}_{j,k}$ analytically, the decision of including the edge (j,k) in the graph could be based on checking whether $\mathfrak{D}_{j,k} > 0$. However, this is usually unrealistic. Therefore a threshold need to be introduced and the decision rule is modified: Include the edge (j,k) into the graph, if

$$\hat{\mathfrak{D}}_{j,k} / \hat{D} > \delta, \quad (32)$$

where $\hat{\mathfrak{D}}_{j,k}$ and \hat{D} are Monte Carlo estimates for D and $\mathfrak{D}_{j,k}$. The threshold δ should be chosen small, e.g. $\delta = 0.01$.

Estimating the measure \mathfrak{D}_{jk} based on Monte Carlo methods is done in the following way: n_{MC} uniform random numbers of $X_{-(jk)}$ are drawn. For each of the resulting points $x_{-(jk)}^{(1)}, \dots, x_{-(jk)}^{(n_1)}$, the function $f_{x_{-(j,k)}^{(i)}}(x_j, x_k)$ is decomposed according to equation (28) and the Sobol index $D_{jk}(x_{-(j,k)}^{(i)})$ is calculated. As this is just a two-dimensional function, the interaction term can be calculated from just knowing the Sobol main effect and the total effect for one of the two parameters. Hence efficient methods like the Extended Fourier Amplitude Sensitivity Test ([14], implemented in the R-package `sensitivity`) can be used for estimating $D_{jk}(x_{-(j,k)}^{(i)})$. Finally the estimate for \mathfrak{D}_{jk} is

$$\hat{\mathfrak{D}}_{jk} = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{D}_{jk}(x_{-(j,k)}^{(i)}). \quad (33)$$

Although applying Monte Carlo methods can be time-consuming, this approach only requires $\binom{d}{2}$ Monte Carlo estimates in contrast to $2^d - 1$ for a conventional, complete Sobol decomposition.

Finally, note that the indices D_j and $\mathfrak{D}_{j,k}$ can be used to provide a quantitative information to the graph. A weight is now added to each edge, proportionally to the value of $\mathfrak{D}_{j,k}$, which indicates the strength of the interaction between the variables x_j and x_k . The same is done for the vertices (materialized by circles), indicating the strength of main effects. Examples are given in the next section.

3.2 Kriging model estimation

The estimation of Gaussian processes defined by equation (21) has been intensively studied when the kernel is a tensor product of 1-dimensional kernels (eq. 23), see e.g. [7], [15] or [11]. Our purpose is to give some insights about its adaptation to kernels associated to cliques defined by eq. (26). First, even if not often encountered in practice, remark that estimation is not always possible with such kernels. Indeed, as remarked by [4] for additive kernels, some special design configurations result in non-invertible covariance matrices. To face this problem, another kernel can be added to the additive decomposition, depending on the whole vector \mathbf{h} :

$$k(\mathbf{h}) = \sum_{l=1}^L k_{C_l}(\mathbf{h}_{C_l}) + k_0(\mathbf{h}) \quad (34)$$

In geostatistics (see e.g. [2]), a common choice for $k_0(\mathbf{h})$ is $\tau^2 \delta_0(\mathbf{h})$, where τ^2 is called a nugget effect, resulting in adding a small positive number on the covariance matrix diagonal. Another choice, also common in geostatistics, is to choose an isotropic kernel, e.g. a kernel defined by equation (eq. 23) with the constraint: $\theta_1 = \dots = \theta_d$. In practice, such problems are not often encountered, because design points are usually chosen at random, and in our experience a small nugget effect is usually enough to overcome the difficulty. Concerning estimation, we have focused on the maximum likelihood estimation (MLE), since it is known to be a good competitor among other existing techniques such as cross-validation. The likelihood expression is recalled in appendix. Its optimization requires special care, due to the problem dimensionality and the multimodalities observed with few data. Three procedures have been implemented: a sequential one as in [7], a direct optimization, and a constrained optimization performed after a change of variables. We refer to the appendix for a precise description. The first one is more time consuming, due to the sequentiality, and has not shown a clear superiority. On the examples presented below, the three algorithms have given comparable results, where the fastest computation times have been achieved by the constrained optimization.

3.3 Global estimation procedure

In the setting of computer experiments there are normally not enough runs for performing Monte Carlo estimates directly on the objective function. As a way out, a first standard Kriging model is estimated. Then the FANOVA graph is estimated from the Kriging interpolator, in replacement of f , as described in 3.1. This may be efficient, provided that the initial Kriging model has enough predictive power. In the following, the clique decomposition of the graph is used to build a kernel structure, and the corresponding model is estimated as shown in 3.2. The global procedure is illustrated in table 1.

Step	
1	For a data set with observations $y^{(1)}, \dots, y^{(n)}, x^{(1)}, \dots, x^{(n)}$, fit a Kriging model with anisotropic product kernel.
2	Use the standard Kriging model for estimating the indices \mathfrak{D}_{jk} .
3	Include the edge (j, k) into the set of edges, E , if $\hat{\mathfrak{D}}_{jk}$ is larger than a tolerance δ .
4	Use the standard Kriging model for estimating the Sobol main effects.
5	Plot the graph using the information from steps 3 and 4.
6	Set up the modified Kriging model with the covariance kernel specified according to the clique structure of the graph.

Table 1: The estimation procedure.

Function name	$f(x)$	Design space
Ishigami	$\sin(x_1) + A\sin^2(x_2) + Bx_3^4\sin(x_1)$	$[-\pi, \pi]^3$
a	$\cos([1, x_1, x_2, x_3]\beta) + \sin([1, x_4, x_5, x_6]\gamma)$	$[-1, 1]^6$
b	$\cos([1, x_1, x_2, x_3]\beta) + \sin([1, x_4, x_5, x_6]\gamma) + ([1, x_3, x_4]\delta)^2$	$[-1, 1]^6$

Table 2: Three analytical examples.

4 Applications to prediction

The design space for the following case studies was always scaled to be $[-1, 1]^d$. Furthermore the settings for the Kriging model as described in section 2.3 and 3.2 are used with a uniform integration measure over the design space. Three analytical examples and two real data sets are considered. The quality of the prediction is judged based on the root mean square error (RMSE).

4.1 Analytical examples

The three analytical examples are shown in table 2. In figure 3 the true graphs for each function are plotted. The following constants are chosen: $A = 7, B = 0.1$ for the Ishigami function as in the introductory section, $\beta = [-0.8, -1.1, 1.1, 1]'$, $\gamma = [-0.5, 0.9, 1, -1.1]'$ and $\delta = [0.5, 0.35, -0.6]'$ for the second and third example.

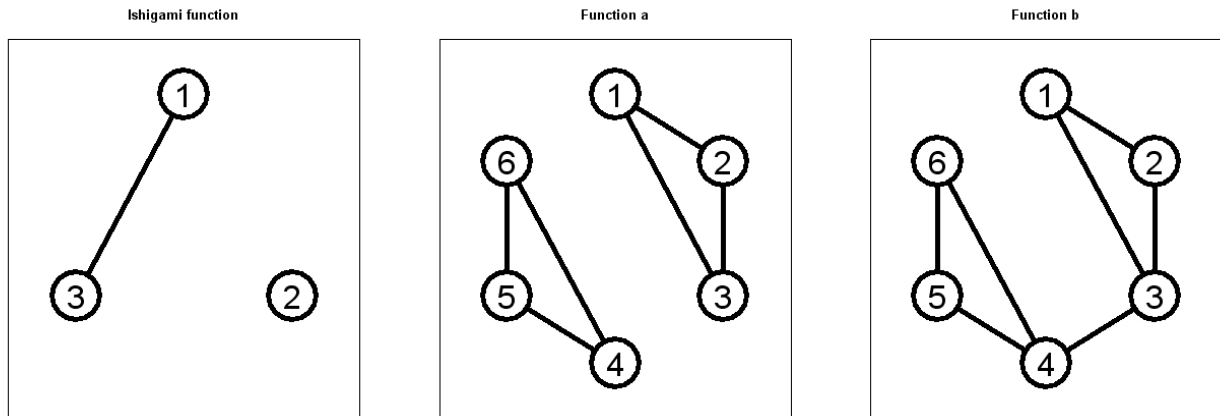


Figure 3: True (unweighted) FANOVA graphs for the analytical examples of table 2.

4.1.1 Ishigami function

The Ishigami function is revised here in order to demonstrate that the process of estimating the graph works properly. Therefore the same setting as in the introduction is used. The standard Kriging model is used as a replacement for the true function and hence all Monte Carlo estimates $\mathfrak{D}_{j,k}, D_j, D$ are based on the standard Kriging model. The resulting graph is shown in figure 4. Here $\delta = 0.1$ was used as a threshold. Figure 4 shows that the graph is estimated correctly. As a comparison, a standard plot for the main effects and total effects based on Sobol indices are plotted on the right hand side. Although in this special case it would be possible to derive the interaction structure from the right hand side plot, this is not the case in general.

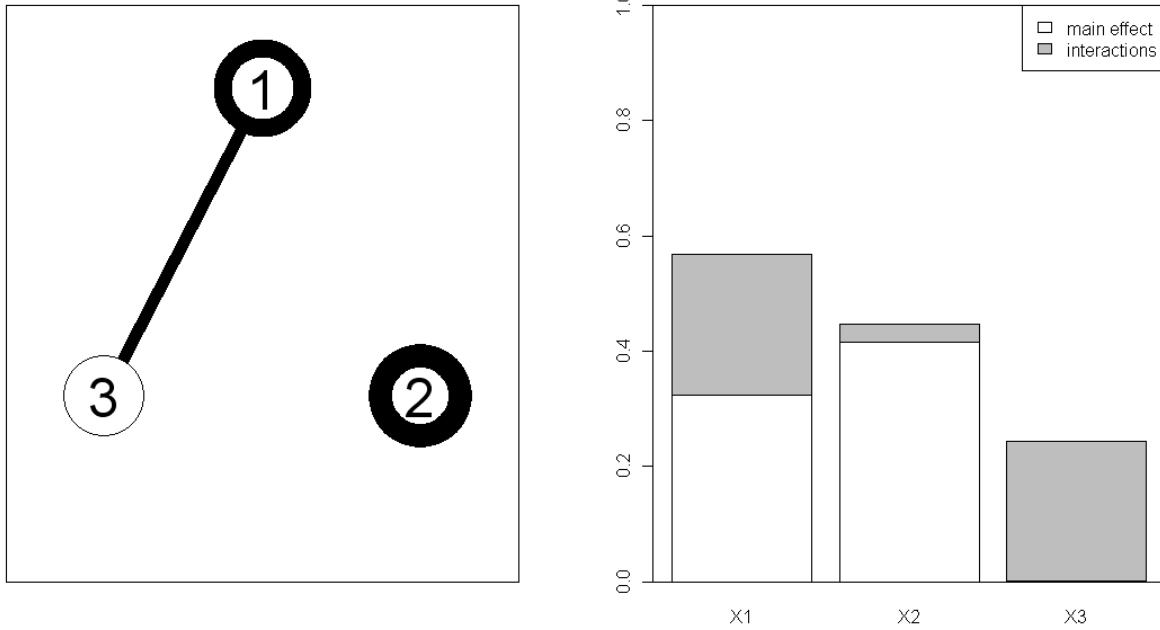


Figure 4: Estimated graph (left hand plot) and Sobol indices (Main effects and total effects, right hand plot) for the Ishigami function

4.1.2 Function a

The Function a is chosen as an example as it has high interaction terms and an interaction structure, which is not just visible from the main/total effects plot. The graph is separable, i.e. the function can be split up into two parts, which are not related to each other. Here (as well as for Function b) 100 runs from a Maximin Latin Hypercube are used for modeling the output and additional 1000 runs from a uniform distribution over the design space are used for comparing the different predictions. The (estimated) graph in figure 5 shows, that the interactions of the clique $\{4, 5, 6\}$ are weaker than that ones for the clique $\{1, 2, 3\}$. The prediction plots in figure 6 show that a big increase of precision is achieved with the modified Kriging model.

4.1.3 Function b

Function b has a close relationship to function a but is not separable. It is visible from the main/total effects plot in figure, that there are differences between function a and function b but it is unclear, what are the core differences with the interactions of function a . In order to clarify the structure of function b again the procedure for estimating the graph is applied, which results in the left hand graph of figure 7. In figure 8, the corresponding prediction plots are drawn. The modified Kriging model delivers a much better prediction than the conventional Kriging model.

Function b can also be used for illustrating strategies for data with high dimensional input space. In such situations it often occurs, that only some input variables have a high influence and many others are of negligible influence. Consider e.g. Function b as a function with 16 input variables, where the first six input variables are chosen as before, and the remaining ten input variables have no influence. Just deleting these input variables is not an option, as these input variables still might have a small effect in reality. The corresponding graph has the same three cliques as before ($C_1 = \{1, 2, 3\}, C_2 = \{4, 5, 6\}, C_3 = \{3, 4\}$) but also ten more cliques for the nonactive input variables ($C_4 = \{7\}, \dots, C_{13} = \{16\}$). Defining the Kriging model according to this clique structure yields a model with very many covariance parameters, which dramatically increase computation times for estimating the parameters as well as reduces prediction power. An alternative is to include all nonactive input variables into one clique with isotropic structure resulting in the cliques $C_1 = \{1, 2, 3\}, C_2 = \{4, 5, 6\}, C_3 = \{3, 4\}, C_4 = \{7, \dots, 16\}$, where C_4 is modeled as the isotropic clique. This decreases the number of covariance parameters from 31 to 13. Comparing three different models (standard Kriging approach, graph based and a graph based kernel with summarizing unimportant variables into an isotropic clique) results figure 9. Here 160 observations according to a Maximin Latin Hypercube have been used for modeling and 1000 additional runs for judging the prediction quality. Here the prediction plots for 1000 additional runs of all three models show, that the modified Kriging model with one isotropic clique for all nonactive input variables performs best with a RMSE of 0.02642 compared to 0.0436(modified Kriging model without isotropic clique) and 0.2301 (standard Kriging model). At the same time it had a much smaller computation time for estimating the covariance parameters than the modified Kriging model with 13 cliques

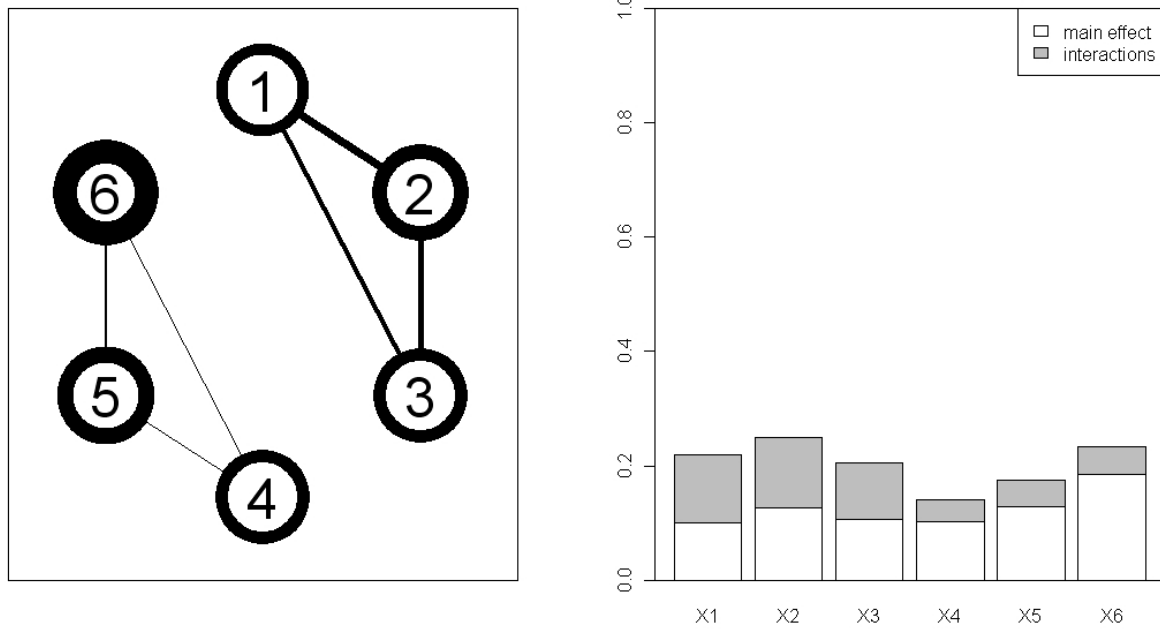


Figure 5: Estimated graph (left hand plot) and Sobol indices (Main effects and total effects, right hand plot) for Function a

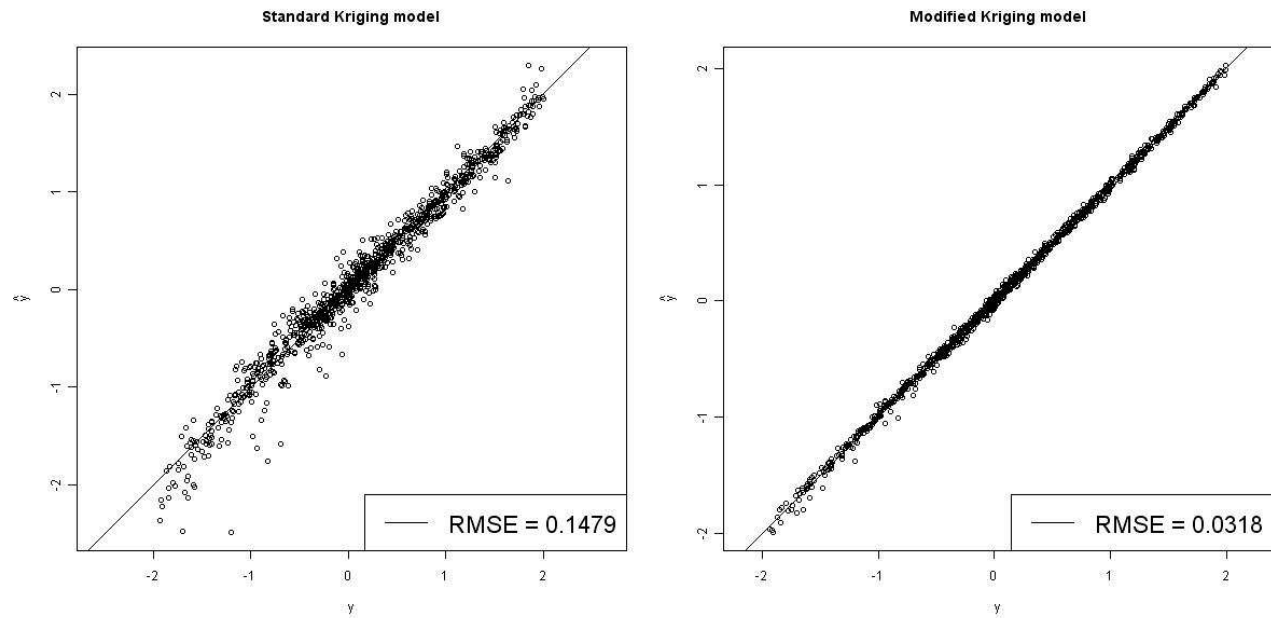


Figure 6: Prediction plots for Function a . On the left hand side for a standard Kriging model, on the right hand side for a modified Kriging model with covariance structure according to figure 5.

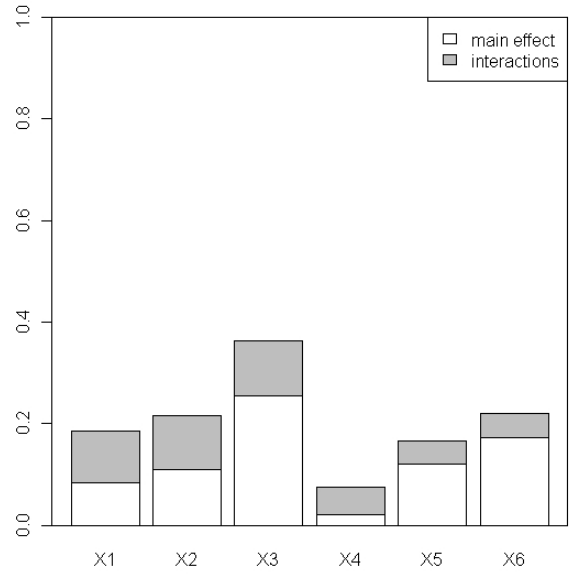
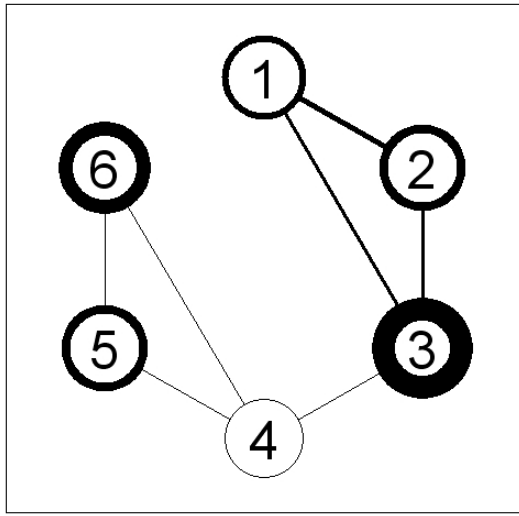


Figure 7: The left hand graph is the estimated one for Function b . On the right hand side Sobol indices (Main effects and total effects) for Function b are plotted.

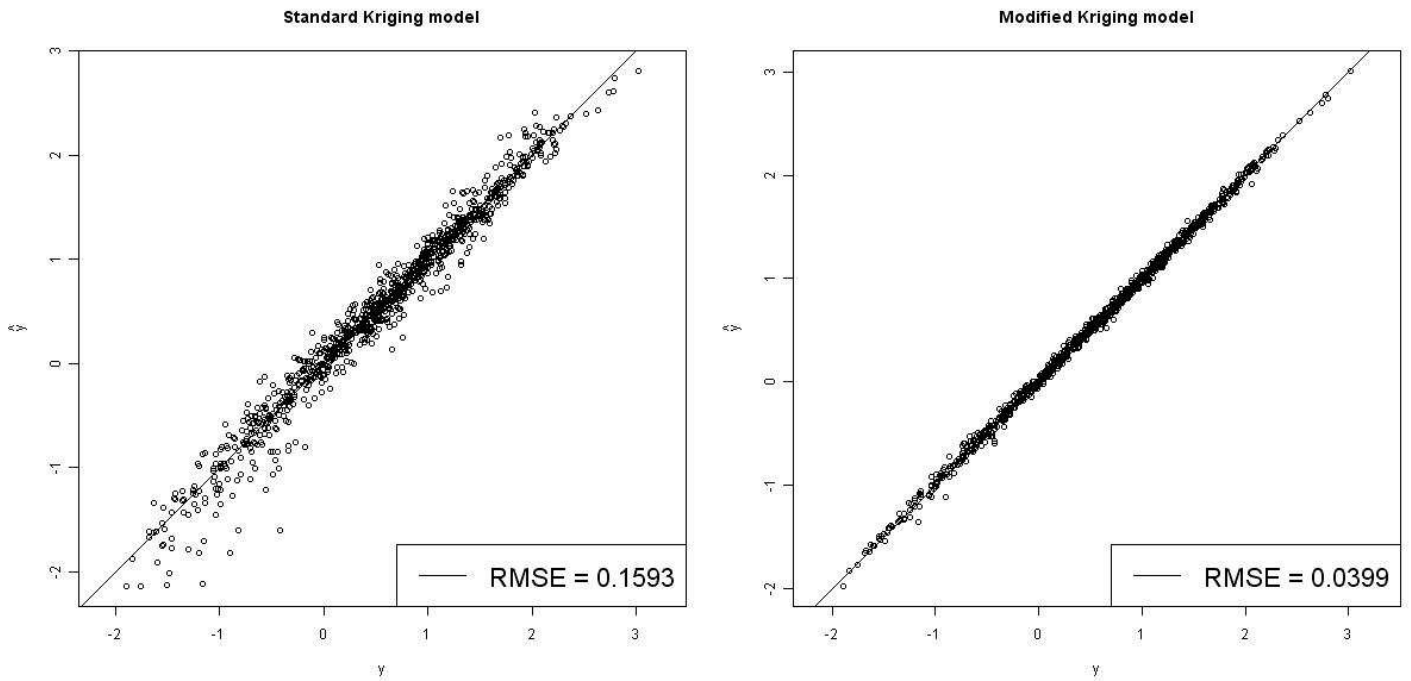


Figure 8: Prediction plots for Function b . On the left hand side for a standard Kriging model and on the right hand side using the modified Kriging model.

(290 seconds compared to 645 seconds.) Hence this can be a strategy for working with high dimensional data situations.

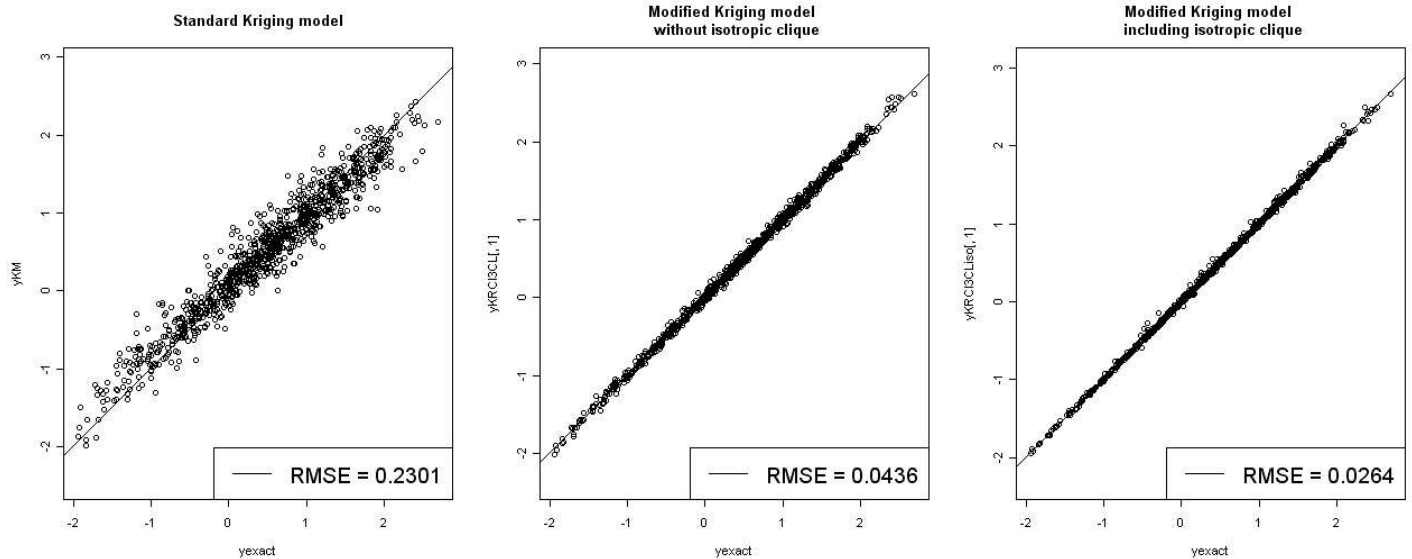


Figure 9: Prediction plots for Function b including 10 inactive factors. On the left hand side for a standard Kriging model, in the middle for a modified Kriging model without an isotropic clique and on the right hand a modified Kriging model including an isotropic clique for inactive factors.

4.2 Case studies

4.2.1 Autoform

The first example is based on data from a case study for the springback analysis during sheet metal forming [8]. The output reflects the amount of spring back after sheet metal forming. The process has been simulated by the engineering software Autoform, which simulates sheet metal forming. The input space is 3-dimensional and there is a 3^3 full factorial design available as learning data set and another 101 runs for validation purposes. Applying the methodology for estimating the graph results in a graph with just on edge $(1,3)$, which can be interpreted such that the second influence parameter just has additive influence. The results for the RMSE and the graphical results show that there is an increase in precision of about 20%.

4.2.2 Piston slap data set

Piston slap is an unwanted noise of engines, which decreases consumer satisfaction. It can be simulated using finite element methods and is analyzed in [7], p. 153 ff. The piston slap can be simulated in dependence of 6 input parameters describing details of the piston. In [7] there are two data set considered. A smaller one with 12 runs and a larger one with 100 runs. [7] use the smaller one for fitting a meta model and the larger one for judging the prediction precision. As the prediction result based on the smaller data set is not satisfying we will for illustration purposes uses the larger data set for fitting a model and the smaller one for validating the prediction results. The results of our analysis, which yield a complex interaction structure, justifies this procedure as the function is too complex for modeling it just based on 12 observations. The resulting graph shows a rather complex interaction structure. Input variable x_6 has interactions with three other input variable whereas for example input variable x_1 , which has the highest main effect, is only interacting with x_6 . Input parameter x_5 shows interesting behavior as it only is active via interactions. Input variable x_3 seems to be completely inactive, as it has no interactions and a very small estimate for the main effect. The resulting graph has 5 cliques: $C_1 = \{1,6\}, C_2 = \{5,6\}, C_3 = \{4,6\}, C_4 = \{2,5\}, C_6 = \{3\}$. Constructing a Kriging model with a correlation structure according this cliques structure, the prediction error can be calculated based on the data set of size 12. The prediction plots for the modified model as well for the standard Kriging model can be found as well in figure 11. The RMSE for the standard Kriging model is 0.1925, whereas for the modified Kriging model a RMSE of 0.1105 is achieved, which represents a substantial improvement for prediction. As the validation data set is of rather smaller size, the RMSE was also estimated by leave-one-out statistics, where for $i = 1, \dots, n$, observation $x^{(i)}, y^{(i)}$ was deleted from the data set and a prediction based on the remaining observations using the parameter estimates for the complete data set was conducted. This yields a RMSE of 0.0864 for the standard Kriging model and a RMSE of 0.0371 for the modified data set, which confirms the improvement of the modified Kriging model.

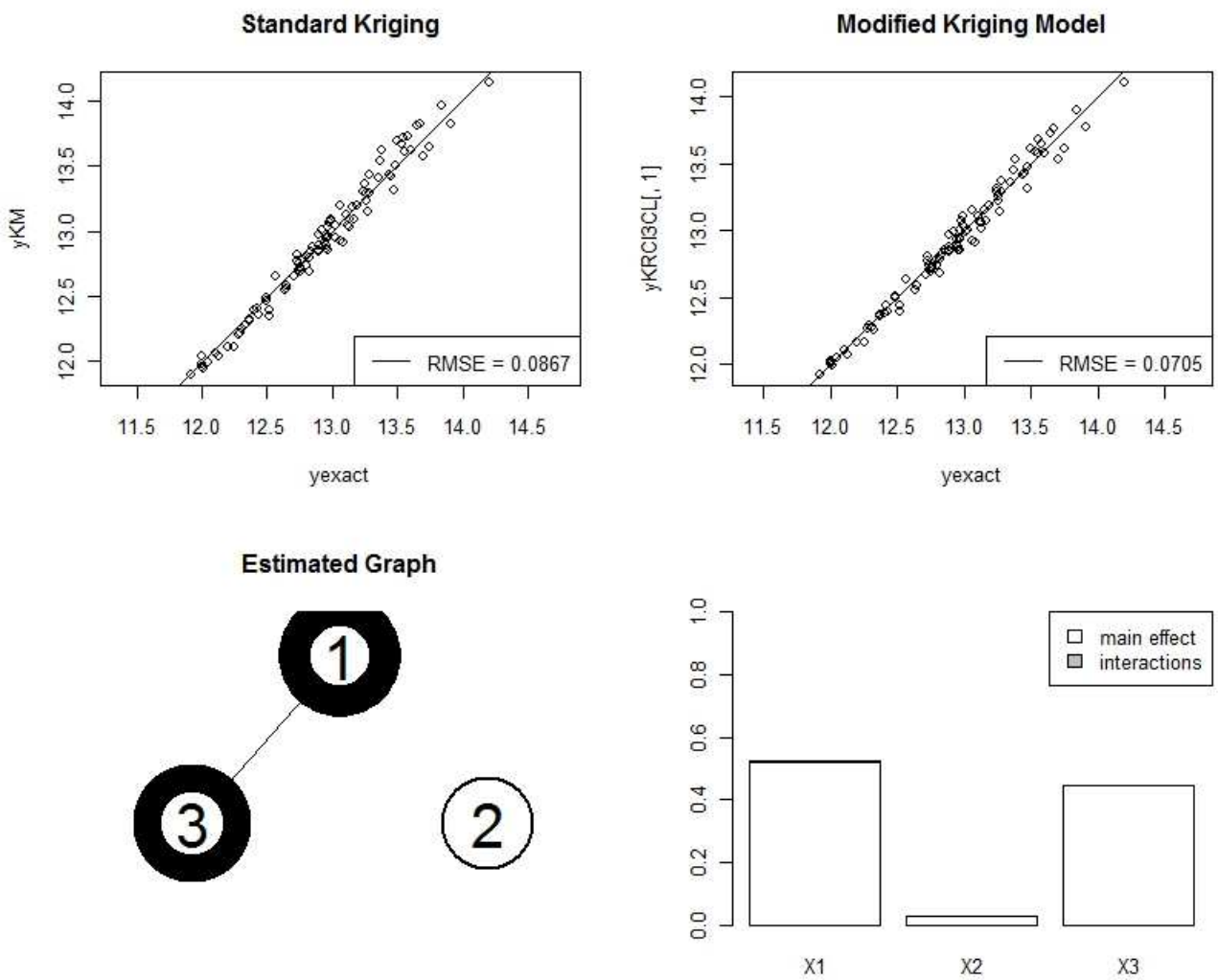


Figure 10: Results for the Autoform data set

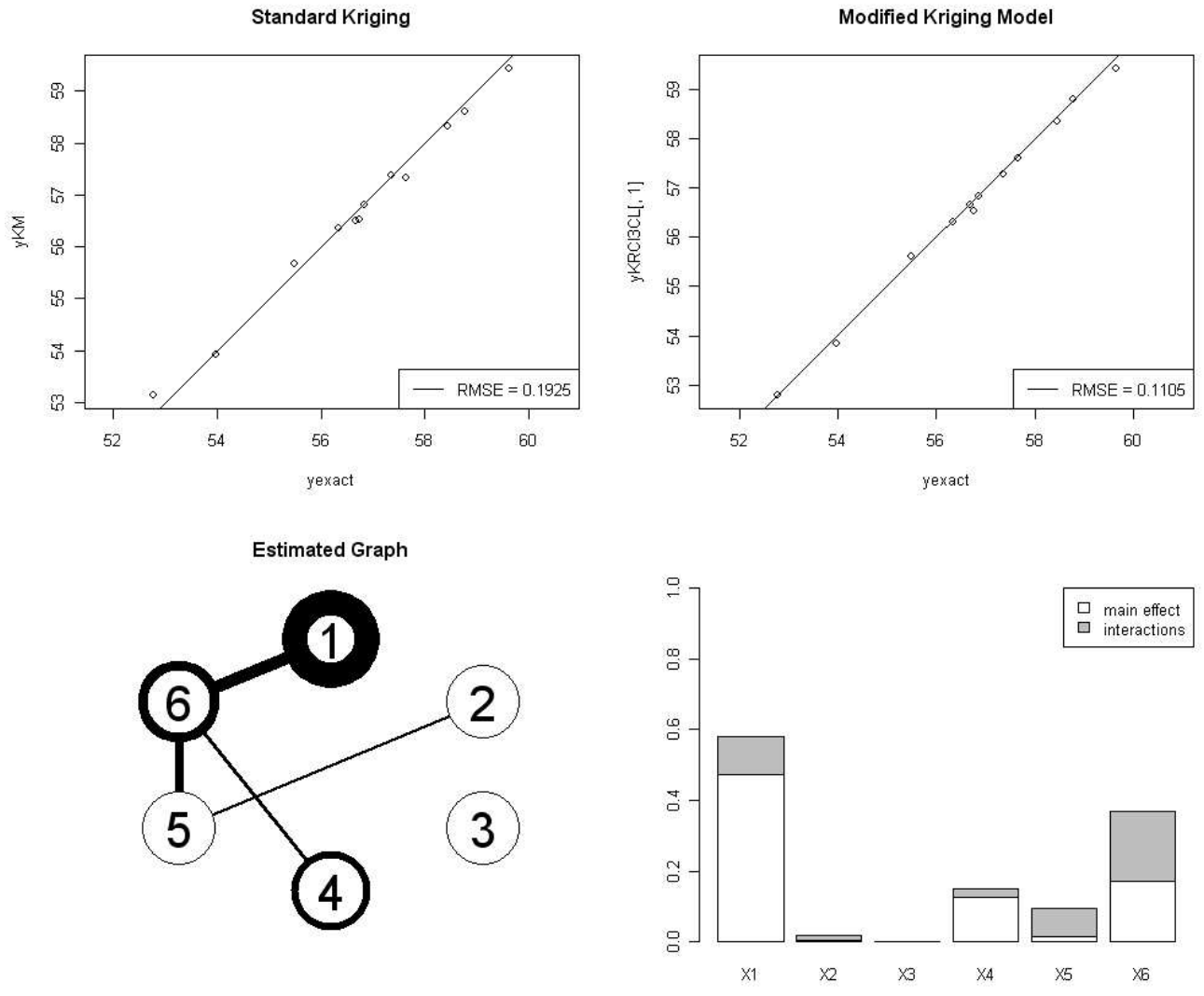


Figure 11: Results for the piston slap data set

5 Discussion

We expect the graph based Kriging prediction to be better than a standard one, if the unknown function has a relatively high degree of complexity and if there are mainly lower dimensional interactions like twofold or threefold active. If the function has a very low complexity (e.g. is mainly linear) than also a standard Kriging model will be able to give a very precise fit and hence the graph based model will not necessarily improve, even eventually decrease prediction power as there are more parameters to be estimated. On the other hand if the unknown function f has the d -fold interaction highly active, the corresponding graph is a complete graph and hence the graph based Kriging model and the standard one are the same.

One of the reasons why the prediction will be better if an adequate graph is chosen, which is not complete is that the resulting prediction function is the sum over several prediction functions, with each prediction function has input dimension lower than d . Hence some kind of dimension reduction is done via replacing a d -dimensional function by a sum over several lower-dimensional functions.

One drawback of the methodology presented here is that a first Kriging model is necessary for estimating the graph. This works only if the first Kriging model has at least some prediction power. Here ways for estimating the graph without requiring a first model are desirable. This can include constructing special designs, which allow for efficient estimation of the graph and sequential methods, which reduce the influence of the initial Kriging model. The definition of the FANOVA graph yields a unique graph, which is then used for defining appropriate covariance kernels. However, in situations with uncertainty a decision between different graphs, which can be reasonable, has to be made. A complex graph can produce overfitting and potentially has very many parameters to be estimated. Hence choosing a good graph in terms of prediction and a simple but adequate model is desirable.

Given the information of the graph $G_{f,v}$ for f , there are several aspects where the graph can be useful besides prediction. One such topic are derivatives, and, subsequently, optimization methods using derivatives. Due to the additive structure of $f(x)$ described by its graph, there is structural information available for the gradient. If the gradient is not known analytically for optimization methods, it normally is estimated. If there is knowledge about the structure of the gradient, this can be incorporated in the algorithm in order to improve convergence.

For robust design problems, where x_1, \dots, x_d are divided into control factors x_K and noise factors x_N , $K \cap N = \emptyset, K \cup N = \{1, \dots, d\}$ (Taguchi situation), the graph can help to illustrate which control factors can be used for controlling the mean of some output and which control factors can be used for controlling the variance of the output.

Other aspects where the graph can be of interest are sensitivity analysis based on Sobol indices and design of experiments. The estimation of Sobol indices based on the modified Kriging prediction function has the potential to perform better than estimates directly obtained from a standard Kriging model as the model already takes into account the interaction structure of the unknown function f . For design of experiments, if the graph belonging to a function f is known a priori, this information can be used to construct space filling designs which are customized to the function f .

6 Conclusion

In this article a methodology has been presented in order to tailor Kriging models to the analysed data set by constructing covariance kernels which take into account the interaction structure of the data generating mechanism. The interaction structure explored by FANOVA methods is presented in a graph which gives an easy to understand graphical illustration of the interaction structure. The clique structure of the resulting graph is then used for defining the data-driven covariance kernels. As shown in section 2.2, the clique structure represents additive parts of the data generating mechanism. Applying this methodology to a data set can result in substantial prediction improvements, especially for modeling functions which have high interactions active but at the same time there are not all input variables interacting with each other. One potential drawback of the method is that in order to estimate the graph and hence the covariance kernel a first conventional (anisotropic) Kriging model has to be constructed. Thus further research will address this issue by considering sequential strategies as well as strategies for defining appropriate graphs and clique structures.

Appendix: MLE for kernels defined by cliques

Suppose that the function f was evaluated at n design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, and denote the vector of observations $\mathbf{y} = (y^{(1)}, \dots, y^{(n)})'$. Assuming that the data are drawn from model (21), \mathbf{y} is normal with mean $\mathbf{F}\boldsymbol{\beta}$ and covariance matrix \mathbf{K} , where :

- $\mathbf{F} = (\mathbf{f}(x^{(1)})', \dots, \mathbf{f}(x^{(n)})')'$, is the $n \times p$ experimental matrix
- $\mathbf{K} = (K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}))_{1 \leq i, j \leq n}$, is the covariance matrix at design points

The likelihood is given by:

$$L(\mathbf{y}; \Psi) = \frac{1}{(2\pi)^{n/2} |\mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})' \mathbf{K}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})\right) \quad (35)$$

Its vector argument Ψ depends on the trend and kernel parameters. Assuming the additive clique decomposition (26), we have:

$$\mathbf{K} = \mathbf{K}_{C_1} + \dots + \mathbf{K}_{C_L}$$

where each \mathbf{K}_{C_l} is the covariance matrix of Z_{C_l} at design points. Due to the stationarity assumption, we have $\mathbf{K}_{C_l} = \sigma_l^2 \mathbf{R}_{C_l}$, where the so-called correlation matrix \mathbf{R}_{C_l} does not depend on σ_l . Thus $\Psi = (\boldsymbol{\beta}, \mathbf{v}, \Theta)'$, where $\mathbf{v} = (\sigma_1^2, \dots, \sigma_L^2)'$ contains the variances parameters and $\Theta = (\Theta_{C_1}, \dots, \Theta_{C_L})'$ the covariances parameters, where each Θ_{C_l} is the vector of the covariance parameters of the kernel k_{C_l} ($l = 1, \dots, L$).

Writing the first order condition results in an analytical expression for $\boldsymbol{\beta}$, as a function of \mathbf{v} and Θ :

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

Therefore maximizing the likelihood (35) is equivalent to maximizing over \mathbf{v} and Θ the "concentrated" log-likelihood obtained by plugging in the expressions of $\hat{\boldsymbol{\beta}}$:

$$-2 \log L(\mathbf{y}; \hat{\boldsymbol{\beta}}, \mathbf{v}, \Theta) = n \log(2\pi) + \log |\mathbf{K}| + (\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})' \mathbf{K}^{-1}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}) \quad (36)$$

Direct optimization. A direct solution consists of optimizing directly (36), using a standard optimization procedure. Note that the analytical gradient can be computed analytically (see e.g. [10] or [11]):

$$-2 \frac{\partial \log L(\mathbf{y}; \hat{\boldsymbol{\beta}}, \mathbf{v}, \Theta)}{\partial \bullet} = -(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})' \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \bullet} \mathbf{K}^{-1}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}) + \text{tr} \left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \bullet} \right)$$

with $\frac{\partial \mathbf{K}}{\partial \theta_{C_l}} = \frac{\partial \mathbf{K}_{C_l}}{\partial \theta_{C_l}}$ and $\frac{\partial \mathbf{K}}{\partial \sigma_l^2} = \mathbf{R}_{C_l}$. It can be given to the optimizer to improve efficiency.

Constrained optimization A drawback of the direct optimization is that it involves unbounded variances parameters, resulting in a huge optimization domain. To overcome this difficulty, in a similar way as [12], the problem can be rewritten using the proportion of variances explained by each clique. Namely, define:

- $\sigma^2 = \sum_{j=1}^L \sigma_j^2$, the total variance
- $\alpha_l = \frac{\sigma_l^2}{\sigma^2}$, the proportion of variance explained by $Z_{C_l}(\cdot)$, $l = 1, \dots, L$

Note that the α_l belong to a L -dimensional simplex with finite volume. Then we have $\mathbf{K} = \sigma^2 \times \mathbf{K}_\alpha$, with $\mathbf{K}_\alpha = \alpha_1 \mathbf{K}_{C_1} + \dots + \alpha_L \mathbf{K}_{C_L}$, and the negative log-likelihood becomes:

$$-2 \ln(L) = n \ln(2\pi) + n \ln(\sigma^2) + \ln |\mathbf{K}_\alpha| + \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{F}\boldsymbol{\beta})' \mathbf{K}_\alpha^{-1} (\mathbf{y} - \mathbf{F}\boldsymbol{\beta})$$

Now writing the first order conditions results in analytical expressions both for $\boldsymbol{\beta}$ and σ^2 (depending on α and Θ):

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{K}_\alpha^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{K}_\alpha^{-1}\mathbf{y} \quad \hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})' \mathbf{K}_\alpha^{-1} (\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})$$

Finally, the problem reduces to optimizing the concentrated log-likelihood

$$-2 \ln L(\mathbf{y}; \hat{\boldsymbol{\beta}}, \hat{\sigma}, \alpha, \Theta) = n \ln(2\pi) + n \ln(\hat{\sigma}^2) + \ln |\mathbf{K}_\alpha| + n$$

over $\alpha \in [0, 1]^k$ and Θ , constrained by $\alpha_1 + \dots + \alpha_L = 1$. The R-command `constrOptim()` can be used for it.

Acknowledgements. The financial support of the DFG (SFB 708, Graduiertenkolleg Statistische Modelbildung) is gratefully acknowledged. Furthermore we would like to thank David Ginsbourger for interesting discussions about this work and many participants of the UCM 2010 conference in Sheffield for many useful comments.

References

- [1] F. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical report, 2009. <http://arxiv.org/abs/0909.0844>.
- [2] N.A.C. Cressie. Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics, 1993.
- [3] R. Diestel. Graph Theory. Springer, New York, 2000.
- [4] N. Durrande, D. Ginsbourger, and O. Roustant. Additive kernels for high-dimensional gaussian process modeling. Technical report, 2010. <http://hal.archives-ouvertes.fr/hal-00446520/en/>.
- [5] D. Edwards. Introduction to Graphical Modelling. Springer-Verlag, New York, 2nd edition, 2000.
- [6] B. Efron and C. Stein. The jackknife estimate of variance. The Annals of Statistics, 9(3):586–596, 1981.
- [7] K.-T. Fang, R. Li, and A. Sudjianto. Design and Modeling for Computer Experiments. Computer Science and Data Analysis Series. Chapman & Hall/CRC, Boca Raton, 2006.
- [8] M. Gsling, H. Kracker, A. Brosius, S. Kuhnt, and A.E. Tekkaya. Simulation und kompensation rckfederungsbedingter formabweichungen. In W. Tillmann, editor, SFB 708 - 3. ffentliches Kolloquium, pages 155 –170. Verlag Praxiswissen, Dortmund, 2009.
- [9] J.E. Oakley and A. O’Hagan. Probabilistic sensitivity analysis of complex models: A bayesian approach. Journal if the Royal Statistical Society B, 66(3):751–769, 2004.
- [10] J.S. Park and J. Baek. Efficient computation of maximum likelihood estimators in a spatial linear model with power exponential covariogram. Computer Geosciences, 27:1–7, 2001.
- [11] C.E. Rasmussen and C.K.I. Williams. Gaussian processes for machine learning. Adaptive computation and machine learning. The MIT Press, Cambridge, 2006.
- [12] O. Roustant, D. Ginsbourger, and Y. Deville. Dicekriging, Diceoptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Technical report, 2010. <http://hal.archives-ouvertes.fr/hal-00495766/fr/>.
- [13] A. Saltelli, K. Chan, and E.M. Scott. Sensitivity Analysis. Wiley Series in Probability and Statistics. Wiley, Chichester, 2000.
- [14] A. Saltelli, Tarantola. S., and K. Chan. A quantitative, model independent method for global sensitivity analysis of model output. Technometrics, 41:39 – 56, 1999.
- [15] T.J. Santner, B.J. Williams, and W.I. Notz. The Design and Analysis of Computer Experiments. Springer Series in Statistics. Springer Verlag, New York, 2003.
- [16] I.M. Sobol’. Sensitivity estimates for nonlinear mathematical models. Mathematical Modeling & Computational Experiment, 1(4):407–414, 1993.
- [17] M.L. Stein. Interpolation of Spatial Data, Some Theory for Kriging. Springer Series in Statistics. Springer, New York, 1999.