ANNALES
DE LA FACULTÉ
DES SCIENCES
DE
TOULOUSE
Mathématiques

NICOLAS DURRANDE, DAVID GINSBOURGER, OLIVIER ROUSTANT

Additive Covariance kernels for high-dimensional Gaussian Process modeling

Tome XXI, n° 3 (2012), p. 481-499.

<http://afst.cedram.org/item?id=AFST_2012_6_21_3_481_0>

Annales de la Faculté des Sciences de Toulouse Vol. XXI, n° 3, 2012
pp. 481–499

Additive Covariance kernels for high-dimensional Gaussian Process modeling

Nicolas Durrande[(1)], David Ginsbourger[(2)],
Olivier Roustant[(3)]

Abstract. — Gaussian Process models are often used for predicting and approximating expensive experiments. However, the number of observations required for building such models may become unrealistic when the input dimension increases. In oder to avoid the curse of dimensionality, a popular approach in multivariate smoothing is to make simplifying assumptions like additivity. The ambition of the present work is to give an insight into a family of covariance kernels that allows combining the features of Gaussian Process modeling with the advantages of generalized additive models, and to describe some properties of the resulting models.

Résumé. — La modélisation par processus gaussiens – aussi appelée krigeage – est souvent utilisée pour obtenir une approximation mathémathique d'une fonction dont l'évaluation est coûteuse. Cependant, le nombre d'évaluations nécessaires pour construire un modèle peut devenir démesuré lorsque la dimension du domaine de définition augmente. Afin de contourner le fléau de la dimension, une alternative bien connue est de se tourner vers des modèles simplifiés comme les modèles additifs. Nous présentons ici une famille de noyaux de covariance permettant de combiner les caractéristiques des modèles de krigeage et les avantages des modèles additifs puis nous décrivons certaines propriétés des modèles obtenus.

(*) Reçu le 28/11/2011, accepté le 10/04/2012

(1) School of mathematics and statistics, University of Sheffield, Sheffield S3 7RH, UK, Ecole Nationale Supérieure des Mines, FAYOL-EMSE, LSTI, F-42023 Saint-Etienne, France
n.durrande@sheffield.ac.uk
(2) Institute of Mathematical Statistics and Actuarial Science, University of Berne, Alpeneggstrasse 22, 3012 Bern, Switzerland
david.ginsbourger@stat.unibe.ch
(3) Ecole Nationale Supérieure des Mines, FAYOL-EMSE, LSTI, F-42023 Saint-Etienne, France
roustant@emse.fr

1. Introduction

High-fidelity numerical simulation studies typically involve calculation intensive computer codes, which underlying costs often imply a drastically limited number of calls to the numerical simulator. Thus, directly coupling a simulator with uncertainty propagation, sensitivity analysis, or global optimization methods is often unaffordable. A well-known approach to circumvent time limitations is to replace the numerical simulator by a mathematical approximation called metamodel (but also emulator, response surface or surrogate model) based on the responses of the simulator for a limited number of inputs called the Design of Experiments (DoE). There are serval families of metamodels, among which the most popular ones include regression, splines, neural networks, and Kriging. In this article, we focus on Kriging, also more recently referred to as Gaussian Process modeling [16]. Originally presented in spatial statistics [5] as an optimal linear unbiased predictor of square integrable random processes, Kriging has become very popular in machine learning, where its interpretation is usually restricted to the convenient framework of Gaussian Processes (GP). The latter point of view allows the explicit derivation of conditional probability distributions for the response values at any point or set of points in the input space.

Since Kriging is usually based on local basis functions, it requires an exponentially increasing number of design points to cover the input space $D \subset \mathbb{R}^d$ when the dimension d increases [19, 6]. A popular approach in multivariate smoothing to get around this issue is to make simplifying assumptions for this case, the emulator m can be decomposed as a sum of univariate functions:

$$m(x) = \mu + \sum_{i=1}^{d} m_i(x_i), \tag{1.1}$$

where $\mu \in \mathbb{R}$ and the m_i's may be non-linear. Since their introduction by Stones in 1985 [21], many methods have been proposed for the estimation of additive models. We can cite the method of marginal integration [15], and a very popular procedure described by Hastie and Tibshirani in [3, 12]: the GAM backfitting algorithm.

However, whatever the chosen estimation technique, the obtained additive models do not completely share the convenient probabilistic framework of GP modeling, including in particular a simply interpretable prediction variance at any input point, or joint conditional distributions at any set of candidate points. Combining the high-dimensional advantages of additive models with the versatility of GPs is the main goal of the present work. For the study of functions that contain an additive part plus a limited number of interactions, developments can be found in [2, 10, 14].

The first part of this article introduces additive Gaussian Processes, their covariance kernels, and the properties of associated Additive Kriging Models (AKM). The second part focuses on the unsuitability of usual separable kernels (e.g. power exponential and Matérn) for high-dimensional modeling and discusses the issue of choosing between a separable or additive kernel. Finally, AKM is compared with standard Kriging models on a well known test function: the Sobol's g-function [18]. It is shown within the latter example that AKM outperforms standard Kriging and produces similar performances as GAM. Due to its approximation performances and its built-in probabilistic framework, the proposed AKM appears as a serious and promising challenger for high-dimensional modeling.

2. Towards additive Kriging

2.1. Additive random processes

Let us here introduce the mathematical construction of additive GPs. A function $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is said additive whenever it can be written as a sum of the form $f(x) = \sum_{i=1}^{d} f_i(x_i)$, where x_i is the i-th component of the d-dimensional input vector x and the f_i's are arbitrary univariate functions. Let us first consider two independent real-valued Gaussian processes Z_1 and Z_2 defined over the same probability space $(\Omega, \mathcal{F}, P)$ and indexed by $\mathbb{R}$, so that their trajectories $Z_i(\cdot;\omega) : t \in \mathbb{R} \to Z_i(t;\omega)$ are univariate real-valued functions. Let $K_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be their respective covariance kernels and $\mu_1, \mu_2 \in \mathbb{R}$ their means. Then, the process Z defined over $(\Omega, \mathcal{F}, P)$ and indexed by $\mathbb{R}^2$, characterized by

$$\forall \omega \in \Omega \ \forall x \in \mathbb{R}^2 \ Z(x;\omega) = Z_1(x_1;\omega) + Z_2(x_2;\omega) \tag{2.1}$$

clearly has additive paths. Z is a Gaussian Process with mean $\mu = \mu_1 + \mu_2$ and kernel $K(x,y) = K_1(x_1,y_1) + K_2(x_2,y_2)$. In this document, we call additive any kernel of the form $K : (x,y) \in \mathbb{R}^d \times \mathbb{R}^d \to K(x,y) = \sum_{i=1}^{d} K_i(x_i,y_i)$ where the K_i's are symmetric positive-semidefinite (s.p.) kernels over $\mathbb{R} \times \mathbb{R}$. Although not commonly encountered in practice, it is well known that such a combination of s.p. kernels is also a s.p. kernel [16, 8]. Moreover, one can show that the paths of any random process with additive kernel are additive in a certain sense:

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 additive, and such that $\forall x \in D,\ \mathrm{P}(Z_x = A_x) = 1$.*

The proof of this proposition is given in the appendix for $d = 2$. For $d \in N$ the proof follows the same pattern but the notations are more cumbersome. Note that the class of additive processes is actually 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 $Z_1(x_1) + Z_2(x_2)$ is also a Gaussian process with additive paths but its kernel is not additive. However, the term additive process will always refer to GPs with additive kernels in this article.

2.2. Invertibility of covariance matrices

As mentioned in [4] the covariance matrix K of observations of an additive process Z at a design of experiments $\mathcal{X} = \{x^{(1)}, \ldots, x^{(n)}\}$ may not be invertible even if there is no redundant point in $\mathcal{X}$. Indeed, the additivity of Z may introduce linear relationships (that hold almost surely) between the observed values of Z and lead to the singularity of K. Figure 1 shows two examples of designs leading to a linear relationship between the observations. For the left panel, the additivity of Z implies that $Z(x^{(4)}) = Z(x^{(2)}) + Z(x^{(3)}) - Z(x^{(1)})$ a.s., so that there is a linear relationship between the columns of K : $K(x^{(i)}, x^{(2)}) + K(x^{(i)}, x^{(3)}) - K(x^{(i)}, x^{(1)}) - K(x^{(i)}, x^{(4)}) = 0$. Therefore, the matrix is not invertible.

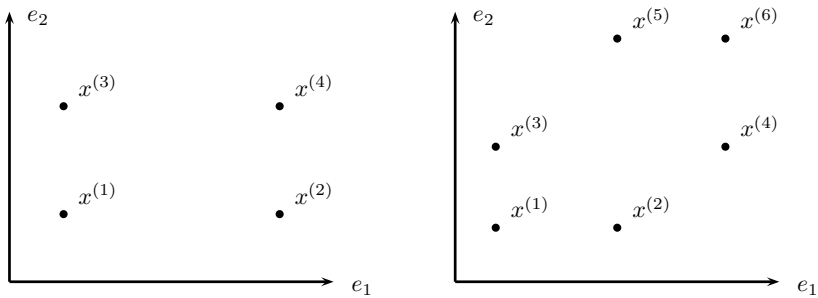


Figure 1. — 2-dimensional examples of DoEs leading to non-invertible covariance matrices when using additive kernels. In both cases, one point can be removed from the DoE without any loss of information.

An approach in accordance with the aim of parsimoniously evaluating the simulator would be to remove some points from the DoE in order to avoid linear combinations. Algebraic methods may be used for determining the subset of points leading to a linear relationship. Indeed, the linear combinations are given by the eigenvectors associated with the null eigenvalues, so the subset of points leading to the non-invertibility of the covariance matrix can be obtained easily. However, the study of a procedure allowing to put aside unnecessary training points is out of the scope of this paper.

An alternative to ensure the invertibility of the covariance matrix is to use a DoE based on Latin Hypercube (LH) sampling. For such designs, the marginals of the DoE $\{x_i^{(1)}, \ldots, x_i^{(n)}\}$ are composed of distinct points. The use of usual stationary kernels implies that the set of random variables $Z_i(x_i^{(1)}),\ \ldots,\ Z_i(x_i^{(n)})$ has a non-degenerate distribution [1, Ex. 3.4] (i.e. the covariance matrix K_i is positive definite). As the sum of positive definite matrices is a still a positive definite matrix, $\mathrm{K} = \sum \mathrm{K}_i$ is then invertible.

2.3. Additive Kriging

Let $f : D \to \mathbb{R}$ be the function of interest (representing for instance the input-output map of a deterministic numerical simulator), where $D \subset \mathbb{R}^d$. The responses of f at the DoE $\mathcal{X}$ are noted $F = (f(x^{(1)})\ \ldots\ f(x^{(n)}))^T$. Simple Kriging relies on the hypothesis that f is one path of a centered square integrable random process Z with kernel K. The formulae for the best predictor and the mean square error (also called *Kriging mean* and *Kriging variance*) are given by:

$$\begin{aligned} m(x) &= \mathrm{E}\left[\, Z(x) |\, Z(\mathcal{X}) = F\right] = k(x)^T \mathrm{K}^{-1} F \\ v(x) &= \mathrm{var}\left[\, Z(x) |\, Z(\mathcal{X}) = F\right] = K(x,x) - k(x)^T \mathrm{K}^{-1} k(x) \end{aligned} \tag{2.2}$$

where $k(\cdot) = \left(K(\cdot, x^{(1)})\ \ldots\ K(\cdot, x^{(n)})\right)^T$ and K is the covariance matrix with entries $\mathrm{K}_{i,j} = K(x^{(i)}, x^{(j)})$. In practice, the structure of K is supposed to be known (e.g. squared-exponential) but its parameters may be unknown. A common way to estimate them is to maximize the likelihood of the parameters given $Z(\mathcal{X}) = F$ [9, 16].

In some cases, the evaluation of f includes an observational noise. Taking this into account in the expression of m and v corresponds to taking the conditional expectation and variance of Z knowing $Z(\mathcal{X}) + \varepsilon = F$. Assuming that ε is a vector of uncorrelated Gaussian variables (white noise) with variance τ^2, we obtain:

$$\begin{aligned} m(x) &= \mathrm{E}\left[\, Z(x) |\, Z(\mathcal{X}) + \varepsilon = F\right] = k(x)^T (\mathrm{K} + \tau^2 \mathrm{Id})^{-1} F \\ v(x) &= \mathrm{var}\left[\, Z(x) |\, Z(\mathcal{X}) + \varepsilon = F\right] = K(x,x) - k(x)^T (\mathrm{K} + \tau^2 \mathrm{Id})^{-1} k(x). \end{aligned} \tag{2.3}$$

The difference with Eq. 2.2 is that the covariance matrix of $\varepsilon(\mathcal{X})$ is added to K in the expression of m and v. As we will use later, this remark is still valid when ε is a correlated Gaussian vector.

Equations 2.2 and 2.3 hold for any s.p. kernel, so they can be applied with additive kernels. In this case, the additivity of the kernel implies the additivity of the Kriging mean so m can be split into a sum of univariate

submodels $m_1, \dots, m_d$. For example, in dimension 2, with a kernel $K(x,y) = K_1(x_1,y_1) + K_2(x_2,y_2)$, we have

$$\begin{aligned} m(x) &= (k_1(x_1) + k_2(x_2))^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} F \\ &= k_1(x_1)^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} F + k_2(x_2)^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} F \\ &= m_1(x_1) + m_2(x_2). \end{aligned} \tag{2.4}$$

Another interesting property concerns the variance: v can be null at points that do not belong to the DoE. Let us consider a two dimensional example where the DoE is composed of the 3 points represented on the left panel of Figure 1: $\mathcal{X} = \{x^{(1)}, x^{(2)}, x^{(3)}\}$. A direct calculation (see Appendix B) shows that the prediction variance at the point $x^{(4)}$ is equal to 0. This particularity follows from the fact that given the observations at $\mathcal{X}$, the value of the additive process at $x^{(4)}$ is known almost surely. In the next section, we illustrate the potential of AKM on an a toy example.

2.4. Illustration and further considerations on a 2D example

We present here a first basic example of an additive Kriging model. We consider $D = [0,1]^2$, and a set of 5 points in D where the value of observations are arbitrarily chosen. Figure 2 shows the obtained Kriging model. We can see on this figure the properties mentioned above: the Kriging mean is an additive function and the prediction variance can be null for points that do no belong to the DoE.

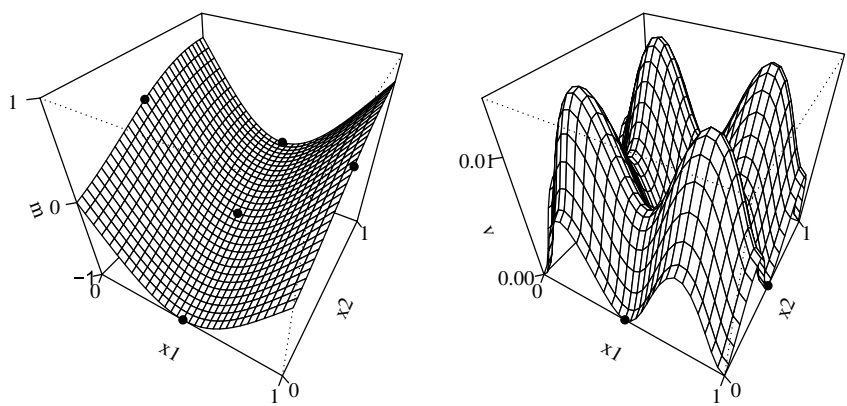


Figure 2. — Approximation by additive Kriging, based on five observations (black dots). The left panel represents the best predictor and the right panel the prediction variance. The kernel here is the additive squared-exponential kernel with parameters $\sigma = (1\ 1)$ and $\theta = (0.6\ 0.6)$.

As seen in Eq. 2.4, the expression of the first univariate model is

$$m_1(x_1) = k_1(x_1)^T(\mathrm{K}_1 + \mathrm{K}_2)^{-1}F. \tag{2.5}$$

Since the matrix K_2 is added to K_1, the contribution of direction 2 appears as observation noise in m_1. We thus get the following expression for the prediction variance

$$v_1(x_1) = K_1(x_1, x_1) - k_1(x_1)^T(\mathrm{K}_1 + \mathrm{K}_2)^{-1}k_1(x_1). \tag{2.6}$$

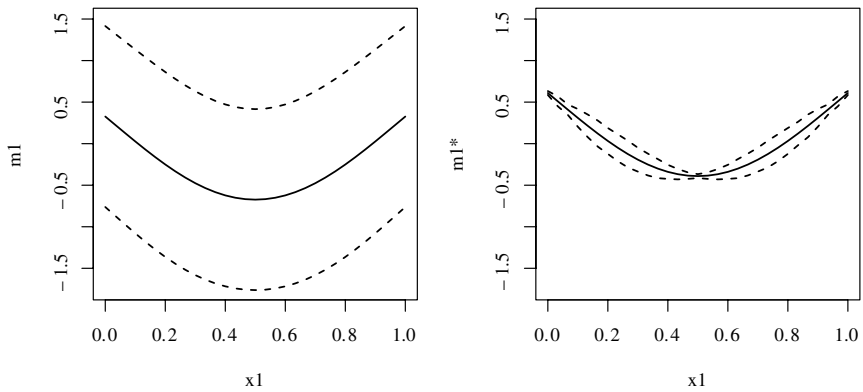


Figure 3. — Univariate models for the 2-dimensional example. The left panel plots m_1 and the 95% confidence intervals $c_1(x_1) = m_1(x_1) \pm 2\sqrt{v_1(x_1)}$. The right panel shows the submodel of the centered univariate effect $\tilde{m}_1$ and $\tilde{c}_1(x_1) = \tilde{m}_1(x_1) \pm 2\sqrt{\tilde{v}_1(x_1)}$.

The left panel of Figure 3 shows the submodel m_1 and the associated 95% confidence intervals. However, it appears that the confidence intervals are wide. This is because the submodels are defined up to a constant. If we assume that $\int Z_i(s_i)\mathrm{d}s_i$ exist a.s. [7], we can get rid of the effect of such a translation by emulating $Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i$ conditionally on the observations:

$$\begin{aligned} \tilde{m}_i(x_i) &= \mathrm{E}\left[Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i \middle| Z(\mathcal{X}) = F \right] \\ \tilde{v}_i(x_i) &= \mathrm{var}\left[Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i \middle| Z(\mathcal{X}) = F \right] \end{aligned} \tag{2.7}$$

The expression of $\tilde{m}_i(x_i)$ is straightforward whereas $\tilde{v}_i(x_i)$ requires more calculations, given in Appendix C.

$$\begin{aligned}\tilde{m}_i(x_i) &= m_i(x_i) - \int m_i(s_i)\mathrm{d}s_i \\ \tilde{v}_i(x_i) &= v_i(x_i) - 2\int K_i(x_i,s_i)\mathrm{d}s_i + 2\int k_i(x_i)^T \mathrm{K}^{-1} k_i(s_i)\mathrm{d}s_i \\ &\quad + K_i(s_i,t_i)\mathrm{d}s_i\mathrm{d}t_i - k_i(t_i)^T\mathrm{K}^{-1}k_i(s_i)\mathrm{d}s_i\mathrm{d}t_i\end{aligned} \tag{2.8}$$

The benefits of using $\tilde{m}_i$ and $\tilde{v}_i$, and to define the submodels up to a constant can be seen on the right panel of Figure 3. Furthermore, as the submodels $\tilde{m}_i$ are univariate and centered, they may give a good approximation of the main effects of the objective function, with relevant confidence intervals. In the end, the probabilistic framework gives an insight on the error for the whole metamodel, but also for each submodel.

3. Kriging, high-dimensional input space and linear budget

We will see in this section that additive Kriging models can outperform usual Kriging models when the dimension of the input space becomes large. The notion of high-dimensional input space can be interpreted differently depending on the context. In our case, we will consider that an input space is high-dimensional when its dimension is larger than 10 and we will consider examples in dimension up to 50. This excludes simulators for which one of the inputs is a picture or a map (e.g., groundwater flow simulators depending on permeability and porosity maps), and for which it is not unusual to deal with 50000-dimensional input spaces.

Most of the time, kernels used in computer experiment are power exponential or Matérn kernels [16]. For those kernels and for all other stationary kernels such that $\lim_{||x-y||\to+\infty} K(x,y) = 0$, an observation at a point $x^{(i)}$ of the DoE has only a local influence on the emulator. This implies that the number of points required for modeling accurately a function increases exponentially with the dimension d of the input space. However, large training sets are rather inconsistent with the context of emulating costly-to-evaluate functions and, in contrast, a common total budget is rather of the order of magnitude of $10 \times d$ evaluations, as advocated in [13].

In the next example, we illustrate that usual separable kernels are not appropriate for emulating high-dimensional functions based on such a linear budget, while additive kernels can be used advantageously to extract the additive trend.

Let Z be a centered Gaussian Process over $[0,1]^d$ with unit variance ($\sigma^2 = 1$) and an isotropic squared-exponential kernel

$$K(x,y) = \sigma^2 \prod_{i=1}^{d} \exp\left(-\frac{(x_i - y_i)^2}{\theta^2}\right). \tag{3.1}$$

Let $\mathcal{X}$ be a LH design of size $10 \times d$. Our aim here is to investigate the evolution of the approximation's quality obtained when conditioning Z on the observations at $\mathcal{X}$, when d increases. In order to quantify the proportion of variance explained by the emulator, we consider a test set $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(n_t)}\}$ drawn from uniform distribution and we compute the following criterion

$$P_{K,\mathcal{X}} = 1 - \frac{\sum_{i=1}^{n_t} \mathrm{var}\left(Z(y^{(i)})|Z(\mathcal{X})\right)}{\sum_{i=1}^{n_t} \mathrm{var}\left(Z(y^{(i)})\right)} = \frac{\sum_{i=1}^{n_t} k(y^{(i)})^T \mathrm{K}^{-1} k(y^{(i)})}{\sum_{i=1}^{n_t} K(y^{(i)}, y^{(i)})}. \tag{3.2}$$

The values of $P_{K,\mathcal{X}}$ are in $[0,1]$ and, as for a Q_2 criterion (see Eq. 5.4), a value $P_{K,\mathcal{X}} = 1$ implies that $Z(y^{(i)})$ is known a.s. for all test points whereas $P_{K,\mathcal{X}} = 0$ indicates that $\mathrm{E}\left(Z(\cdot)|Z(\mathcal{X})\right)$ is no more predictive than $\mathrm{E}\left(Z(\cdot)\right)$. As it is based on a the IMSE, $P_{K,\mathcal{X}}$ quantifies the reduction of variance due to the knowledge of $Z(\mathcal{X})$ so it asses if a model is a priori predictive or not. However, it is not meant to quantify the accuracy the approximation of f by m and v.

As shown on Figure 4, the proportion of explained variance collapses when the dimension increases, and this fall is all the more important as the range parameter θ is small. For $d > 15$ and $\theta < 0.5$ (i.e. θ is lower than half of the length of the marginals of $[0,1]^d$), a budget of $10 \times d$ observations is not sufficient for Kriging models based on usual separable covariance. However, further tests showed that for $\theta = \sqrt{d}$ such budget allows to build very predictive GP emulator up to $d = 100$.

We will now consider a second example where the GP to be approximated has an explicit additive component and compare the results of additive and classical Kriging emulators. Let Y_A and Y_S be independant centered GPs indexed by $[0,1]^d$ with respectively an additive and a separable kernel:

$$\begin{aligned} K_A(x,y) &= \frac{1}{d} \sum_{i=1}^{d} \exp\left(-\frac{(x_i - y_i)^2}{0.5^2}\right) \\ K_S(x,y) &= \prod_{i=1}^{d} \exp\left(-\frac{(x_i - y_i)^2}{0.5^2}\right). \end{aligned} \tag{3.3}$$

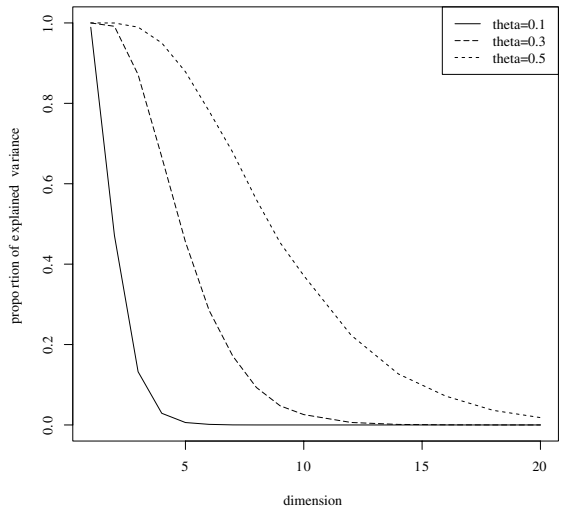


Figure 4. — Proportion of variance explained by the knowledge of $Z(\mathcal{X})$ versus dimension. The $P_{K,\mathcal{X}}$ criteria is computed for $n_t = 10000$ test points uniformly distributed on $[0,1]^d$. The 3 curves correspond to different values of the range parameter θ, as detailed on the figure.

We define Y as $Y = Y_A + Y_S$ so that the first half of the variance of Y is explained by its additive part Z_A and the second one by its separable part Z_S. We now compare the predictivity of 2 emulators:

$$\begin{aligned} m_A(x) &= \mathrm{E}(Y_A(x)|Y_A(\mathcal{X}) + Y_S(\mathcal{X})) \\ &= k_A(x)^t(\mathrm{K}_A + \mathrm{K}_S)^{-1}(Y_A(\mathcal{X}) + Y_S(\mathcal{X})) \\ m_S(x) &= \mathrm{E}(Y_S(x)|Y_A(\mathcal{X}) + Y_S(\mathcal{X})) \\ &= k_S(x)^t(\mathrm{K}_A + \mathrm{K}_S)^{-1}(Y_A(\mathcal{X}) + Y_S(\mathcal{X})). \end{aligned} \tag{3.4}$$

As we have seen previously, m_A corresponds to the best predictor of an additive Kriging model with an observation noise given by K_S. This emulator cannot explain the non additive part of Y. Reciprocally, m_S is based on the separable kernel K_S with an observation noise K_A. This second model can potentially cover both the additive and non additive part of Y for a large number of observations.

The prediction variance associated with those emulators is known analytically, so $P_{K_A,\mathcal{X}}$ and $P_{K_S,\mathcal{X}}$ can be compared as in the previous example. We observe on Figure 5 that the explained variance falls quickly to 0 when using a separable kernel whereas an emulator based on an additive kernel can capture efficiently the additive part of the phenomena. In this example,

it appears that for a budget of $10 \times d$ evaluations, additive Kriging models clearly outperform Kriging based on standard kernels.

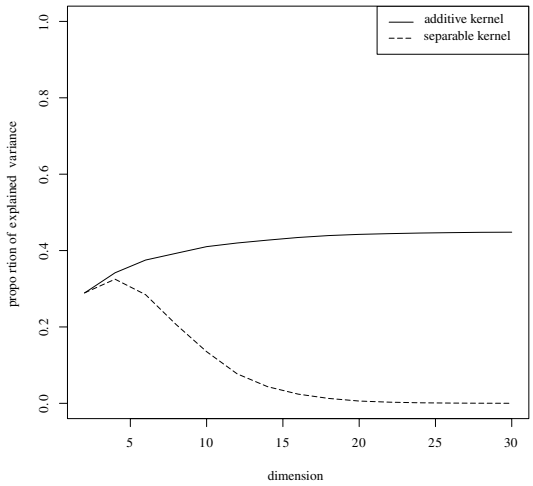


Figure 5. — Comparison of the predictivity of the approximation of $Y = Y_A + Y_S$ by m_A and m_S.

4. The issue of choosing a kernel

Choosing between an additive and a tensor product kernel is an important issue. In this section, we give some guidelines based on the predictivity of the two types of models. Let f be a real-valued function over $[0,1]^d$, assumed square integrable for the uniform measure μ. The ANOVA representation of f is its decomposition as a sum of terms with increasing interaction order

$$f(x) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j}^{d} f_{i,j}(x_{i,j}) + \ldots + f_{1,\ldots,d}(x_{1,\ldots,d}) \tag{4.1}$$

where unicity is guaranteed by side conditions $\forall I \subset \{1,\ldots,d\},\ \forall i \in I$, $\int f_I(x_I)\mathrm{d}x_i = 0$. The terms of Eq. 4.1 can be regrouped to obtain

$$f(x) = f_0 + f_{add}(x) + f_{int}(x) \tag{4.2}$$

where $f_{add} = \sum_{i=1}^{d} f_i(x_i)$ represents the additive part of f and f_{int} stands for all the interaction terms. Following the ANOVA framework [20], we define the additivity ratio as

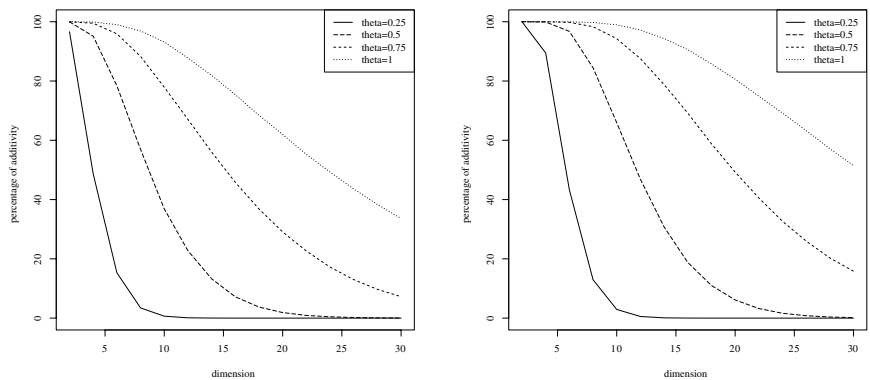
$$q = \frac{\mathrm{var}(f_{add}(X))}{\mathrm{var}(f(X))} \tag{4.3}$$

where X is a random variable with distribution μ.

Answering the question "Given my problem, would it be better to use a tensor product or an additive kernel?" may have huge practical implications. However, since the adequacy on one kernel or the other depends on f, there is no universal answer to that question. Nevertheless, it is still possible to compare the predictivity of the two models (as previously) in order to choose the one with the best prediction ability. We thus reformulate the previous question as "For given univariate kernels K_i, from which value of the additivity ratio q is the additive model more predictive? (or inversely less predictive?)".

For these settings, we consider that the predictivity of the models based on $K_S = \prod K_i$ and $K_A = \sum K_i$ are respectively given by $P_{K_S,\mathcal{X}}$ and $q \times P_{K_A,\mathcal{X}}$. Thus, the two models are equally predictive when $P_{K_S,\mathcal{X}} = q \times P_{K_A,\mathcal{X}}$ so the ratio $q = P_{K_S,\mathcal{X}}/P_{K_A,\mathcal{X}}$ corresponds to the percentage of additivity such that the models have the same prediction ability.

Figure 6 shows $P_{K_S,\mathcal{X}}/P_{K_A,\mathcal{X}}$ as a function of d when K_i are squared exponential kernels. For a given value of the range parameter θ, the points above the curve correspond to values of d and q for which an additive model is more likely to be appropriate whereas a tensor product kernel is more suitable for the region below the curve. The comparison between the two panels shows that increasing the number of points of the DoE benefits more to the separable kernel than to the additive one.



(A) $n = 10 \times d$ (B) $n = 50 \times d$

Figure 6. — Additivity ratio q such that the predictivity of the additive model $q \times P_{K_A,\mathcal{X}}$ is equal to the one of an usual tensor product kernel (i.e. $P_{K_S,\mathcal{X}}$). The models are based on univariate squared exponential kernels with range parameter θ. The actual percentage of predictivity, which is not represented, lies in the region between the curves and the x axis. The comparison between the two panels gives an insight on the influence of the DoE's number of points.

This type of graphics may be used to choose between an additive or a tensor product kernel. However, it is important to recall that a good predictivity criterion does not necessarily implies a good accuracy, and that the estimation of the range parameter is likely to differ depending on the kind of kernel (additive or of tensor product type). As a conclusion to this section, it appears that additive models should be used in low dimension only if the function to approximate has an important additive part. On the other hand, in high dimension additive models can offer better predictions than usual separable models even for functions with a small additivity ratio.

5. 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 function to approximate is the g-function of Sobol defined over $[0,1]^d$ by

$$g(x) = \prod_{k=1}^{d} \frac{|4x_k - 2| + a_k}{1 + a_k} \text{ with } a_k > 0. \tag{5.1}$$

This popular function from the literature [18] is obviously not additive. However, depending on the coefficients a_k, g can be very close to an additive function. As a rule, the g-function is all the more additive as the a_k are large. One main advantage for our study is that the Sobol sensitivity indices can be obtained analytically, so that we can quantify the ratio of additivity of the test function. For $i = 1, \ldots, d$ the index S_i associated with the variable x_i is

$$S_i = \frac{\frac{1}{3(1+a_i)^2}}{\left[\prod_{k=1}^{d} 1 + \frac{1}{3(1+a_k)^2}\right] - 1}. \tag{5.2}$$

Here, we impose that the value of the parameters a_k is the same for all directions (i.e. $\forall k,\ a_k = a_1$). As the additivity of the g-function is tunable, we choose a_1 such that the percentage of additivity of g is 75%:

$$\sum_{i=1}^{d} S_i = 0.75 \Leftrightarrow d\frac{u}{(1+u)^d - 1} = 0.75 \quad \text{with} \quad u = \frac{1}{3(1+a_1)^2}. \tag{5.3}$$

Eventually, the value of a_1 can be obtained by finding the zeros of a polynomial in u. Note that different values for d lead to different values of a_1.

For $d \in \{5, 10, 20, 30\}$ and a Latin hypercube design based on $10 \times d$ points, we compare an Usual Kriging Model (UKM) with both AKM, and a Generalized Additive Model (GAM) obtained with the backfitting algorithm [12]. As the latter is based on smoothing cubic splines, we choose for

the Kriging models a Matérn 5/2 kernel with observation noise to ensure that the different models have a similar regularity. All Kriging models include a constant term as trend, so are they Ordinary Kriging models. The results for UKM and GAM are obtained with the DiceKriging [17] and the GAM [11] *R* packages available on the CRAN website [22]. For AKM and UKM the kernel parameters (σ^2, θ) and the observational noise's variance τ^2 are obtained using maximum likelihood estimation [16, 19]. To asses the accuracy of the obtained metamodels, the Q_2 coefficient is computed on a test sample of $n_t = 1000$ points uniformly distributed over $[0,1]^d$:

$$Q_2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n_t}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.4}$$

where y is the vector of actual response values at the test points, $\hat{y}$ is the vector of predicted values and $\bar{y}$ is the mean of y.

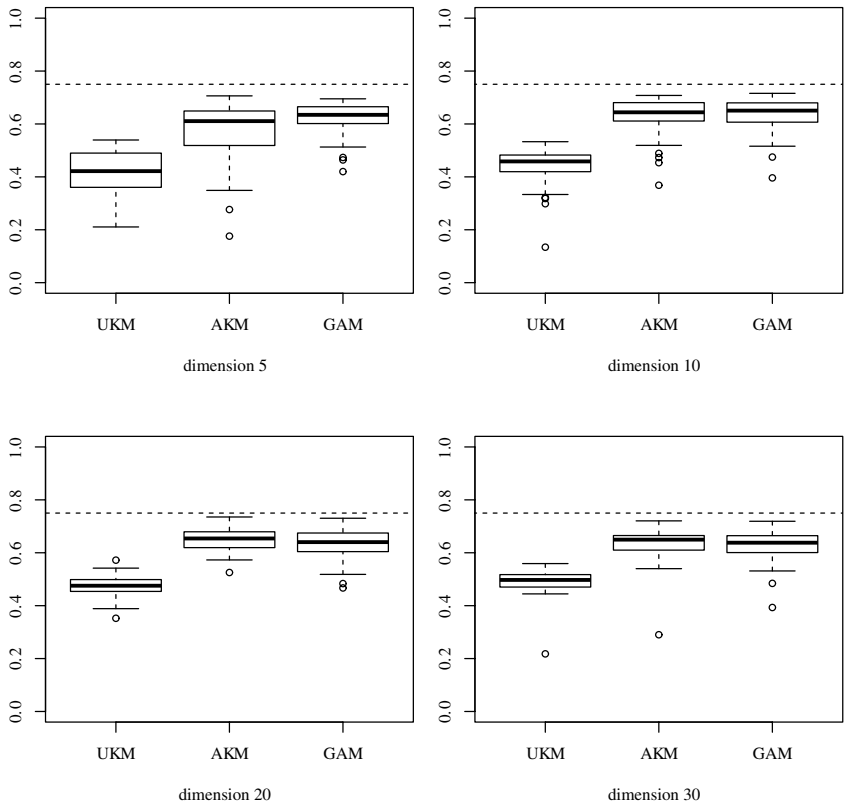
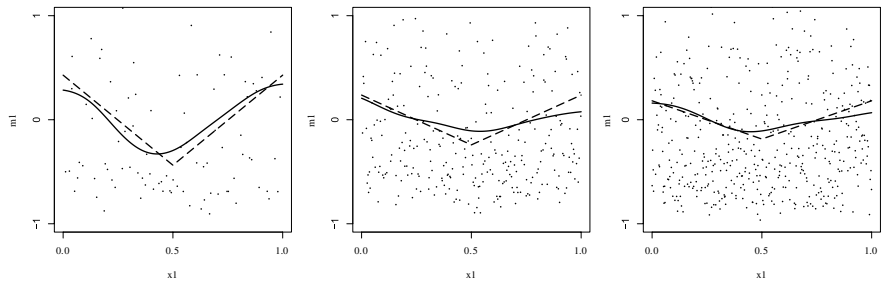


Figure 7. — Boxplots of the Q_2 coefficients for three emulators: Usual Kriging Model (UKM), Additive Kriging Model (AKM) and GAM. For a given boxplot, the variability is due to the choice of the DoE which is repeated 50 times.

As the parameter estimation accuracy and the overall quality of the emulators are likely to fluctuate with the DoE, we repeated 50 times the Q_2 computation for various DoEs. The results are presented in Figure 7. Conversely to what we observed in section 3, the predictivity of the Kriging model based on a separable kernel does not fall to zero when the dimension increases. Indeed, as we impose the additive part of g to explain 75% of its variance, the value of the coefficient a_1 is increasing with d and the g-function becomes smoother. As a result, the range parameter θ increases with d (we have $\theta \approx 0.5$ for $d = 5$ and $\theta \approx 2$ for $d = 30$), explaining the observed effect.

Since one can plot the submodels, the additivity of the best predictors allows to illustrate this increasing smoothness of g. For example, Figure 8 shows that the univariate submodels $\tilde{m}_1$ becomes flatter with increasing d. On these graphics, the submodels are close to the analytical main effects even if the observation points do not show any obvious trend.



(A) $d = 10$ (B) $d = 30$ (C) $d = 50$

Figure 8. — Representation of the univariate submodels $\tilde{m}_1(x_1)$ (solid lines) for three additive Kriging models. As a comparison, the analytical main effects are given by the dashed lines. The bullets denote the centered observation points.

6. Concluding remarks

The proposed methodology seems to be a good challenger for additive modeling. On the first example, additive models appears to be well suited for high-dimensional modeling with a DoE budget of $10 \times d$ whereas Kriging models based on standard kernels fail to recover the function to approximate. One important result is that additive Kriging models succeed to extract the additive trend of the function to approximate even if this function is not purely additive.

In section 4, we discussed briefly the choice between an additive or a separable kernel. This issue is of great importance for the practical use of

Kriging models. It appeared on the example that increasing the dimension or the percentage of additivity favors additive models whereas increasing the number of points in the DoE or the values of the range parameter is in the advantage of models based on tensor product kernels.

The proposed Kriging models benefits from the properties of additive models, while taking advantage from GP features. For the first point we can cite the complexity reduction and the interpretability of additive models. For the second, the main asset is that GP models include a prediction variance for the model but also for each submodel. This justifies the fact of modeling an additive function on $\mathbb{R}^d$ instead of building d metamodels over $\mathbb{R}$ since the prediction variance is not additive. In the end, the proposed methodology is fully compatible with Kriging-based methods and their versatile applications. Potential perspectives include the use of additive and related Kriging models for optimization, e.g. relying on infill sampling criteria like the Expected Improvement.

Note finally that only isotropic kernels were considered in this article. As for separable kernels, the use of additive kernels could easily be extended to anisotropic kernels (i.e. one range parameter θ_i per direction); futhermore, additive kernels also allow to define one variance parameter σ_i^2 per direction. This feature, which is not available for separable kernels, can enable additive models to better approximate functions for which the variance depends on the direction. Of course, the total number of parameters would be $2d + 1$, and the practicability of their estimation deserves to be studied in more detail.

Bibliography

[1] Azaïs (J.M.) and Wschebor (M.). — Level sets and extrema of random processes and fields, Wiley Online Library (2009).

[2] Bach (F.). — Exploring large feature spaces with hierarchical multiple kernel learning, Arxiv preprint arXiv:0809.1493 (2008).

[3] Buja (A.), Hastie (T.) and Tibshirani (R.). — Linear smoothers and additive models, The Annals of Statistics, p. 453-510 (1989).

[4] Chilès (J.P.) and Delfiner (P.). — Geostatistics: modeling spatial uncertainty, volume 344, Wiley-Interscience (1999).

[5] Cressie (N.). — Statistics for spatial data, Terra Nova, 4(5), p. 613-617 (1992).

[6] Fang (K.). — Design and modeling for computer experiments, volume 6. CRC Press (2006).

[7] Fortet (R.M.). — Les operateurs integraux dont le noyau est une covariance, Trabajos de estadística y de investigación operativa, 36(3), p. 133-144 (1985).

[8] Gaetan (C.) and Guyon (X.). — Spatial statistics and modeling, Springer Verlag (2009).

[9] Ginsbourger (D.), Dupuy (D.), Badea (A.), Carraro (L.) and Roustant (O.). — A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments, Applied Stochastic Models in Business and Industry, 25(2), p. 115-131 (2009).

[10] Gunn (S.R.) and Brown (M.). — Supanova: A sparse, transparent modelling approach, In Neural Networks for Signal Processing IX, 1999, Proceedings of the 1999 IEEE Signal Processing Society Workshop, p. 21-30. IEEE (1999).

[11] Hastie (T.). — gam: Generalized Additive Models, 2011, R package version 1.04.1.

[12] Hastie (T.J.) and Tibshirani (R.J.). — Generalized additive models, Chapman & Hall/CRC (1990).

[13] Loeppky (J.L.), Sacks (J.) and Welch (W.J.). — Choosing the sample size of a computer experiment: A practical guide, Technometrics, 51(4), p. 366-376 (2009).

[14] Muehlenstaedt (T.), Roustant (O.), Carraro (L.) and Kuhnt (S.). — Data-driven Kriging models based on FANOVA-decomposition, to appear in Statistics and Computing.

[15] Newey (W.K.). — Kernel estimation of partial means and a general variance estimator, Econometric Theory, 10(02), p. 1-21 (1994).

[16] Rasmussen (C.E.) and Williams (C.K.I.). — Gaussian processes for machine learning (2005).

[17] Roustant (O.), Ginsbourger (D.) and Deville (Y.). — DiceKriging: Kriging methods for computer experiments, 2011, R package version 1.3.

[18] Saltelli (A.), Chan (K.), Scott (E.M.) et al. — Sensitivity analysis, volume 134, Wiley New York (2000).

[19] Santner (T.J.), Williams (B.J.) and Notz (W.). — The design and analysis of computer experiments, Springer Verlag (2003).

[20] Sobol (I.M.). — Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates, Mathematics and Computers in Simulation, 55(1-3), p. 271-280, (2001).

[21] Stone (C.J.). — Additive regression and other nonparametric models, The annals of Statistics, p. 689-705 (1985).

[22] R Team. — R: A language and environment for statistical computing, R Foundation for Statistical Computing Vienna Austria ISBN, 3(10) (2008).

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

Let Z be a centered random process indexed by $\mathbb{R}^2$ with covariance 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 the sake of simplicity, the three terms of $\mathrm{var}[Z(x) - Z_T(x)] = \mathrm{var}[Z(x)] + \mathrm{var}[Z_T(x)] - 2\mathrm{cov}[Z(x), Z_T(x)]$ are studied separately:

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

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

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

Those three equations imply that $\mathrm{var}[Z(x) - Z_T(x)] = 0$, $\forall x \in \mathbb{R}^2$. As $\mathrm{E}[Z(x) - Z_T(x)] = 0$, we have $\mathrm{P}(Z(x) = Z_T(x)) = 1$ so Z_T is a modification of Z with additive paths.

Appendix B: Calculation of the prediction variance

Let consider a DoE composed of the 3 points $\{x^{(1)}, x^{(2)}, x^{(3)}\}$ represented on the left panel of Figure 1. We want here to show that although $x^{(4)}$ does not belongs to the DoE we have $v(x^{(4)}) = 0$.

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

$$
\begin{aligned}
& (-1 \quad 1 \quad 1) \begin{pmatrix} K_1(x_1^{(1)}, x_1^{(4)}) + K_2(x_2^{(1)}, x_2^{(4)}) \\ K_1(x_1^{(2)}, x_1^{(4)}) + K_2(x_2^{(2)}, x_2^{(4)}) \\ K_1(x_1^{(3)}, x_1^{(4)}) + K_2(x_2^{(3)}, x_2^{(4)}) \end{pmatrix} \\
= \; & K_1(x_1^{(2)}, x_1^{(2)}) + K_2(x_2^{(3)}, x_2^{(3)}) - K_1(x_1^{(2)}, x_1^{(2)}) - K_2(x_2^{(3)}, x_2^{(3)}) \\
= \; & 0
\end{aligned}
$$

Appendix C: Calculation of $\tilde{v}_i$

We want here to calculate the variance of $Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i$ conditionally to the observations Y.

$$
\begin{aligned}
\tilde{v}_i(x_i) &= \mathrm{var}\left[Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i \middle| Z(X) = Y \right] \\
&= \mathrm{var}\left[Z_i(x_i) \middle| Z(X) = Y \right] - 2\mathrm{cov}\left[Z_i(x_i), \int Z_i(s_i)\mathrm{d}s_i \middle| Z(X) = Y \right] \\
&\qquad + \mathrm{var}\left[\int Z_i(s_i)\mathrm{d}s_i \middle| Z(X) = Y \right] \\
&= v_i(x_i) - 2\left(\int K_i(x_i, s_i)\mathrm{d}s_i - \int k_i(x_i)^T K^{-1} k_i(s_i)\mathrm{d}s_i \right) \\
&\qquad + \iint K_i(s_i, t_i)\mathrm{d}s_i\mathrm{d}t_i - \iint k_i(t_i)^T K^{-1} k_i(s_i)\mathrm{d}s_i\mathrm{d}t_i.
\end{aligned}
$$