ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis

N. Durrande[a,b,*], D. Ginsbourger[c], O. Roustant[b], L. Carraro[d]

[a] *School of mathematics and statistics, University of Sheffield, Sheffield S3 7RH, UK*
[b] *Ecole Nationale Supérieure des Mines, FAYOL-EMSE, LSTI, F-42023 Saint-Etienne, France*
[c] *Institute of Mathematical Statistics and Actuarial Science, University of Berne, Alpeneggstrasse 22 - 3012 Bern, Switzerland*
[d] *TELECOM St Etienne, 25 rue du Dr Rémy Annino - 42000 Saint Etienne, France*

Abstract

Given a reproducing kernel Hilbert space $(\mathcal{H}, \langle .,. \rangle)$ of real-valued functions and a suitable measure μ over the source space $D \subset \mathbb{R}$, we decompose $\mathcal{H}$ as the sum of a subspace of centered functions for μ and its orthogonal in $\mathcal{H}$. This decomposition leads to a special case of ANOVA kernels, for which the functional ANOVA representation of the best predictor can be elegantly derived, either in an interpolation or regularization framework. The proposed kernels appear to be particularly convenient for analyzing the effect of each (group of) variable(s) and computing sensitivity indices without recursivity.

Keywords: Gaussian process regression, global sensitivity analysis, Hoeffding-Sobol Decomposition, SS-ANOVA

1. Introduction

Let f be a real-valued function defined over $D \subset \mathbb{R}^d$. We assume that f is costly to evaluate and that we want to study some global properties of f such as the influence of a variable x_i on the output $f(x_1, \ldots, x_d)$. As the number of evaluations of f is limited, it may be unaffordable to run sensitivity analysis methods directly on f. Thus, it can be helpful to replace f by a mathematical approximation for performing such studies [12]. In this article we propose a class of models that is well suited for performing global sensitivity analysis. First, we present some background in sensitivity analysis, interpolation in Reproducing Kernel Hilbert Space (RKHS), and on the class of ANOVA kernels. Then we construct RKHS of zero mean functions and derive a subclass of ANOVA kernels that is well suited for the ANOVA representation of the best predictor and for global sensitivity analysis. The use of this new class of kernels is illustrated

*Corresponding author. Nicolas Durrande, email: durrande@emse.fr, phone: (+44) (0)114 222 3824.

throughout the article on classical test functions from the sensitivity analysis literature.

1.1. Sensitivity analysis

The purpose of global sensitivity analysis is to describe the inner structure of f and to analyse the influence of each variable or group of variables on f [15]. A traditional approach is to study the variance of $f(\mathbf{X})$ where $\mathbf{X}$ is a random vector over D. Hereafter, we assume that $f \in L^2(D,\mu)$, where $D = D_1 \times \cdots \times D_d$ is a Cartesian product space of sets $D_i \subset \mathbb{R}$ and where $\mu = \mu_1 \otimes \cdots \otimes \mu_d$ is a product of probability measures over D_i. The measure μ describes the variability of the inputs and we define $\mathbf{X}$ as a random vector with probability distribution μ. Note that although this framework is often considered in sensitivity analysis, the product structure of μ implies that the components of $\mathbf{X}$ are independent.

For $d = 1$, any $g \in L^2(D,\mu)$ can be decomposed as a sum of a constant plus a zero mean function [8],

$$g = \int_D g(s)\mathrm{d}\mu(s) + \left(g - \int_D g(s)\mathrm{d}\mu(s)\right).$$

The two elements of this decomposition are orthogonal for the usual L^2 scalar product so we have a geometric decomposition of $L^2(D,\mu)$:

$$L^2(D,\mu) = L^2_1(D,\mu) \overset{\perp}{\oplus} L^2_0(D,\mu) \tag{1}$$

where $L^2_1(D,\mu)$ denotes the subspace of constant functions and $L^2_0(D,\mu)$ the subspace of zero mean functions: $L^2_0(D,\mu) = \{g \in L^2(D,\mu) : \int_D g(s)\mathrm{d}\mu(s) = 0\}$.

Similarly, if $d > 1$, the space $L^2(D,\mu)$ has a tensor product structure [10]

$$L^2(D,\mu) = \bigotimes_{i=1}^{d} L^2(D_i,\mu_i). \tag{2}$$

Using Eq. 1 and the notation $L^2_B(D,\mu) = \bigotimes_{i=1}^{d} L^2_{B_i}(D_i,\mu_i)$ for $B \in \{0,1\}^d$ we obtain

$$L^2(D,\mu) = \bigotimes_{i=1}^{d} \left(L^2_1(D_i,\mu_i) \overset{\perp}{\oplus} L^2_0(D_i,\mu_i)\right) = \bigoplus_{B\in\{0,1\}^d}^{\perp} L^2_B(D,\mu). \tag{3}$$

A key property is that two subspaces L^2_B and $L^2_{B'}$ are orthogonal whenever $B \neq B'$. Given an arbitrary function $f \in L^2(D,\mu)$, the orthogonal projection of f onto those subspaces leads to the functional ANOVA representation [7, 17] (or *Hoeffding-Sobol decomposition*) of f into main effects and interactions:

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j} f_{i,j}(x_i,x_j) + \cdots + f_{1,\ldots,d}(\mathbf{x}), \tag{4}$$

where f_0 is the orthogonal projection of f onto the space of constant functions $L^2_{\{0\}^d}(D,\mu)$ and where f_I ($I \subset \{1,\dots,d\}$) is the orthogonal projection of f onto $L^2_{\mathrm{ind}(I)}(D,\mu)$ with $\mathrm{ind}(I)_i = 1$ if $i \in I$ and $\mathrm{ind}(I)_i = 0$ if $i \notin I$. By construction, the integral of f_I with respect to any of the variables indexed by $i \in I$ is zero. This decomposition gives an insight into the influence of each set of variables $\mathbf{x}_I = \{x_i, i \in I\}$ on f. For the constant term, the main effects, and the two-factor interactions, one gets the classical expressions [8]:

$$\begin{aligned} f_0 &= \int_D f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\ f_i(x_i) &= \int_{D_{-i}} f(\mathbf{x})\mathrm{d}\mu_{-i}(\mathbf{x}_{-i}) - f_0 \\ f_{i,j}(x_i,x_j) &= \int_{D_{-\{i,j\}}} f(\mathbf{x})\mathrm{d}\mu_{-\{i,j\}}(\mathbf{x}_{-\{i,j\}}) - f_i(x_i) - f_j(x_j) - f_0 \end{aligned} \tag{5}$$

with the notations $D_{-I} = \prod_{i\notin I} D_i$ and $\mu_{-I} = \bigotimes_{i\notin I} \mu_i$. As one can see from Eq. 5, the calculation of any f_I requires having recursively computed all the f_J's for $J \subset I$, which makes it cumbersome (if not practically impossible) to get high order interactions.

Recall the random vector $\mathbf{X}$ with distribution μ. $L^2(D,\mu)$-orthogonality between any two terms of the decomposition implies that the variance of the random variable $f(\mathbf{X})$ can be decomposed as

$$\mathrm{var}(f(\mathbf{X})) = \sum_{i=1}^{d} \mathrm{var}(f_i(X_i)) + \sum_{i<j} \mathrm{var}(f_{i,j}(\mathbf{X}_{i,j})) + \cdots + \mathrm{var}(f_{1,\dots,d}(\mathbf{X})). \tag{6}$$

For the subsets $I \subset \{1,\dots,d\}$, the global sensitivity indices S_I (also called Sobol indices) are then defined as

$$S_I = \frac{\mathrm{var}(f_I(\mathbf{X}_I))}{\mathrm{var}(f(\mathbf{X}))}. \tag{7}$$

S_I represents the proportion of variance of $f(\mathbf{X})$ explained by the interaction between the variables indexed by I. The knowledge of the indices S_I is very helpful for understanding the influence of the inputs, but the computation of the f_I's is cumbersome when evaluation of f is costly since they rely on the computation of the integrals of Eq. 5. Following [12], it can then be advantageous to perform the sensitivity analysis on a surrogate model m approximating f.

1.2. Optimal approximation in RKHS

The class of functional approximation techniques considered in this work, commonly referred to as Kriging or Gaussian process regression in contemporary statistical learning settings, can be seen as methods of optimal interpolation in

RKHS. f is here assumed to be known at a set of points $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ with $\mathcal{X}_i \in D$. Given a RKHS $\mathcal{H}$ of real-valued functions over D with reproducing kernel $K(.,.)$, we want to select a function of $\mathcal{H}$ that is a good approximation of f. Two options are usually considered [13]: the first one is to solve an interpolation problem, and the second a regularization problem.

The interpolation approach consists of finding a function $h \in \mathcal{H}$ that satisfies $h(\mathcal{X}_i) = f(\mathcal{X}_i)$ for $i = 1, \ldots, n$ (for which we will use the vectorial notation $h(\mathcal{X}) = f(\mathcal{X})$). As there might be an infinite number of function of $\mathcal{H}$ that satisfy this criterion, the optimal interpolator is defined as [13]:

$$m(\mathbf{x}) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \left(||h||_{\mathcal{H}} : h(\mathcal{X}) = f(\mathcal{X}) \right). \tag{8}$$

When the Gram matrix K of general term $\mathrm{K}_{i,j} = K(\mathcal{X}_i, \mathcal{X}_j)$ is invertible, we obtain

$$m(\mathbf{x}) = \mathbf{k}(\mathbf{x})^t \mathrm{K}^{-1} \mathbf{F} \tag{9}$$

where $\mathbf{F} = f(\mathcal{X})$ is the column vector of observations and $\mathbf{k}(.)$ is the column vector of functions $(K(\mathcal{X}_i, .))_{1 \leq i \leq n}$.

For the regularization approach, the best predictor $\tilde{m}$ is defined to be the minimizer of:

$$\sum_{i=1}^{n} (h(\mathcal{X}_i) - F_i)^2 + \lambda ||h||^2_{\mathcal{H}}, \tag{10}$$

where the parameter λ is a positive number tuning the trade-off between the norm of $\tilde{m}$ and the distance to the observations. The solution of this minimization problem is

$$\tilde{m}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^t (\mathrm{K} + \lambda \mathrm{I})^{-1} \mathbf{F}. \tag{11}$$

From the probabilistic point of view, the regularization corresponds to approximating f based on noisy observations : $f_{obs}(\mathcal{X}_i) = f(\mathcal{X}_i) + \varepsilon_i$ where the ε_i are independent random variables with distribution $\mathcal{N}(0, \lambda)$ [13].

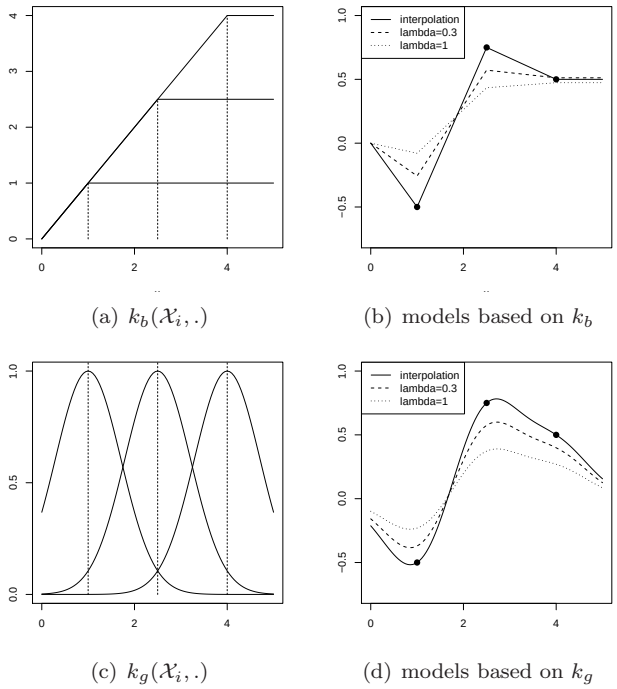
As it appears in Eqs 9 and 11, m and $\tilde{m}$ are a linear combinations of the $K(\mathcal{X}_i, .)$. These basis functions and the influence of the parameter λ are illustrated in Figure 1 on a toy example for two popular kernels:

$$k_b(x, y) = \min(x, y) \qquad \text{and} \qquad k_g(x, y) = \exp\left(-(x - y)^2\right), \tag{12}$$

known respectively as the *Brownian* and the *Gaussian* covariance kernels.

The interpolation problem can be seen as a particular case of a regularization problem where $\lambda \to 0$ [8]. For the sake of clarity, we focus hereafter on the best interpolator m. However the general case of the regularization framework is addressed in Section 4.

The kernel K can be any symmetric positive definite (s.p.d.) function [19] and it has to be chosen in practice. It is customary to arbitrarily select K from usual s.p.d. functions (e.g. Matérn, or power-exponential kernel families)



(a) $k_b(\mathcal{X}_i, .)$ (b) models based on k_b

(c) $k_g(\mathcal{X}_i, .)$ (d) models based on k_g

Figure 1: Examples of models obtained with kernels k_b and k_g with $D = [0, 5]$. The set $\mathcal{X} = \{1, 2.5, 4\}$ and the observations $\mathbf{F} = (-0.5, 0.75, 0.5)^t$ are chosen arbitrarily. The left panels represent the basis functions on which the model is based and the right panels show the obtained models m and $\tilde{m}$ for two values of λ.

according to some prior knowledge about f such as its regularity. However, this choice has great impact on the resulting model (e.g. Figure 1). If a kernel depends on some parameters, a traditional approach is to estimate them by maximum likelihood to ensure the kernel is adapted to the data [13].

In the next section we focus on a particular family of s.p.d. functions called ANOVA kernels, which are designed to offer a separation of the terms with different orders of interaction. The main contribution of this paper, which is presented in the third section, deals with a special case of ANOVA kernels tailored for an improved disentanglement of multivariate effects.

1.3. ANOVA kernels and a candidate ANOVA-like decomposition of the best predictor

The name "ANOVA kernel" has been introduced in the literature of machine learning by Stitson et Al. [18] in the late 90s. These kernels allow control of the order of interaction in a model and enhance their interpretability [9, 4]. For $D = D_1 \times \cdots \times D_d$, they are constructed as a product of univariate kernels $1 + k^i$, where 1 stands for a *bias term* and the k^i's are arbitrary s.p.d. kernels

on $D_i \times D_i$ $(1 \leq i \leq d)$:

$$K_{ANOVA}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{d}(1 + k^i(x_i, y_i)) = 1 + \sum_{I \subset \{1,\dots,d\}} \prod_{i \in I} k^i(x_i, y_i). \tag{13}$$

As shown on the right hand side of Eq. 13, ANOVA kernels can also be seen as a sum of separable kernels with increasing interaction orders [6].

Denoting by $\mathbb{1}^i$ and $\mathcal{H}^i$ the RKHS of functions defined over D_i with respective reproducing kernels 1 and k^i, K_{ANOVA} is the reproducing kernel of the space [2, 4]

$$\mathcal{H}_{ANOVA} = \bigotimes_{i=1}^{d}(\mathbb{1}^i + \mathcal{H}^i) = \mathbb{1} + \sum_{I \subset \{1,\dots,d\}} \mathcal{H}_I. \tag{14}$$

with $\mathcal{H}_I = \bigotimes_{i \in I} \mathcal{H}^i \otimes \bigotimes_{i \notin I} \mathbb{1}^i$.

From Eq. 13 the particular structure of K_{ANOVA} allows us to develop the $n \times 1$ vector $\mathbf{k}(\mathbf{x})$ of Eq. 9 as follows:

$$\mathbf{k}(\mathbf{x}) = \mathbf{1} + \sum_{I \subset \{1,\dots,d\}} \bigodot_{i \in I} \mathbf{k}^i(x_i) \tag{15}$$

where $\odot$ denotes a element-wise product: $\left(\bigodot_{i \in I} \mathbf{k}^i(x_i)\right)_j = \prod_{i \in I} k^i(x_i, x_j)$. From this relation, we can get the decomposition of the best predictor m given in Eq. 9 onto the subspaces $\mathcal{H}_I$:

$$m(\mathbf{x}) = \mathbf{1}^t \mathrm{K}^{-1}\mathbf{F} + \sum_{I \subset \{1,\dots,d\}} \left(\bigodot_{i \in I} \mathbf{k}^i(x_i)\right)^t \mathrm{K}^{-1}\mathbf{F} \tag{16}$$

Noting $m_0 = \mathbf{1}^t \mathrm{K}^{-1}\mathbf{F}$ and $m_I(\mathbf{x}) = \left(\bigodot_{i \in I} \mathbf{k}^i(x_i)\right)^t \mathrm{K}^{-1}\mathbf{F}$, we obtain an expression for m which is similar to its ANOVA representation:

$$m(\mathbf{x}) = m_0 + \sum_{i=1}^{d} m_i(x_i) + \sum_{i<j} m_{i,j}(\mathbf{x}_{i,j}) + \cdots + m_{1,\dots,d}(\mathbf{x}_{1,\dots,d}). \tag{17}$$

In this expression, the m_I have the useful feature of not requiring any recursive computation of integrals. However, this decomposition differs from the ANOVA representation of m since the properties of Eq. 4 are not respected. Indeed, the m_I of Eq. 17 are not necessarily zero mean functions and any two terms of the decomposition are generally not orthogonal in $L^2(D, \mu)$. This point is illustrated numerically on a test function at the end of the second example. It can be seen theoretically if the k_i are Ornstein-Uhlenbeck kernels. In this case, it is known that $\mathbb{1}^i \subset \mathcal{H}^i$ [1]. As all the $\mathcal{H}_I$ contain $\mathbb{1}$ they are obviously not orthogonal.

An alternative to avoid this issue is to consider RKHS such that for all i, $\mathcal{H}^i$

is L^2-orthogonal to the space of constant functions $\mathbb{1}^i$ (i.e. the $\mathcal{H}^i$ are spaces of zero mean functions for μ_i). This construction ensures that the decomposition of Eq. 16 has the properties required in Eq. 4 so we benefit from the advantages of the two equations: the meaning of Eq. 4 for the analysis of variance; and the simplicity of computation for the m_I's from Eq. 16. This approach has been applied to splines under the name of Smoothing Spline ANOVA (SS-ANOVA) [20, 8]. However, SS-ANOVA is based on a decomposition of the inner product of a spline RKHS and it cannot be immediately applied for other RKHS. In the next section we introduce a general method for extracting RKHS of zero mean functions from any usual RKHS. As this approach is not limited to splines, it can be seen as an extension of the SS-ANOVA framework.

2. RKHS of zero mean functions

2.1. Decomposition of one-dimensional RKHS

Let $\mathcal{H}$ be a RKHS of functions defined over a set $D \subset \mathbb{R}$ with kernel k and μ a finite Borel measure over D. Furthermore, we consider the hypothesis:

H 1.
(i) $k : D \times D \to \mathbb{R}$ *is* $\mu \otimes \mu$*-measurable.*
(ii) $\int_D \sqrt{k(s,s)}\mathrm{d}\mu(s) < \infty$.

Note that when is μ a probability measure, any bounded kernel (e.g. Gaussian, power-exponential, Matérn) satisfies the condition (ii). Thus H1 is not a very restrictive hypothesis.

Proposition 1. *Under H1,* $\mathcal{H}$ *can be decomposed as a sum of two orthogonal sub-RKHS,* $\mathcal{H} = \mathcal{H}_0 \overset{\perp}{\oplus} \mathcal{H}_1$ *where* $\mathcal{H}_0$ *is a RKHS of zero-mean functions for* μ*, and its orthogonal* $\mathcal{H}_1$ *is at most 1-dimensional.*

Proof. From H1, the integral operator $I : \mathcal{H} \to \mathbb{R},\ h \mapsto \int_D h(s)\mathrm{d}\mu(s)$ is bounded since

$$|I(h)| \leq \int_D |\langle h, k(s,.)\rangle_{\mathcal{H}}|\mathrm{d}\mu(s) \leq ||h||_{\mathcal{H}} \int_D \sqrt{k(s,s)}\mathrm{d}\mu(s) \tag{18}$$

According to the Riesz representation theorem, there exists a unique $R \in \mathcal{H}$ such that $\forall h \in \mathcal{H},\ I(h) = \langle h, R\rangle_{\mathcal{H}}$. If $R(.) = 0$, then all $f \in \mathcal{H}$ are centered functions for μ, so that $\mathcal{H}_0 = \mathcal{H}$ and $\mathcal{H}_1 = \{0\}$. If $R(.) \neq 0$, then $\mathcal{H}_1 = span(R)$ is a 1-dimensional sub-RKHS of $\mathcal{H}$, and the subspace $\mathcal{H}_0$ of centered functions for μ is characterized by $\mathcal{H}_0 = \mathcal{H}_1^{\perp}$. □

For all $x \in D$ the value of $R(x)$ can be calculated explicitly. Indeed, recalling that $k(x,.)$ and R are respectively the representers in $\mathcal{H}$ of the evaluation functional at x and of the integral operator, we get:

$$R(x) = \langle k(x,.), R\rangle_{\mathcal{H}} = I(k(x,.)) = \int_D k(x,s)\mathrm{d}\mu(s). \tag{19}$$

The reproducing kernels k_0, k_1 of $\mathcal{H}_0$ and $\mathcal{H}_1$ satisfy $k = k_0 + k_1$. Let π denote the orthogonal projection onto $\mathcal{H}_1$. Following [4, theorem 11] we obtain

$$\begin{aligned} k_0(x,y) &= k(x,y) - \pi(k(x,.))(y) \\ &= k(x,y) - \frac{\int_D k(x,s)\mathrm{d}\mu(s) \int_D k(y,s)\mathrm{d}\mu(s)}{\iint_{D\times D} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)} \end{aligned} \tag{20}$$

Example 1. Let us illustrate the previous results for the kernels k_b and k_g introduced in Eq. 12. As previously, we choose $D = [0, 5]$ and define μ as the uniform measure over D. Following Proposition 1, k_b and k_g can be decomposed as a sum of two reproducing kernels:

$$\begin{aligned} k_b(x,y) &= k_{b,0}(x,y) + k_{b,1}(x,y) \\ k_g(x,y) &= k_{g,0}(x,y) + k_{g,1}(x,y). \end{aligned} \tag{21}$$

Figure 2 represents sections of $k_{b,i}(x, y)$ and $k_{g,i}(x, y)$ for various values of y. We observe from this figure that $k_{b,0}(., y)$ and $k_{g,0}(., y)$ are zero mean functions. Moreover, $k_{b,0}(., y)$ and $k_{b,1}(., y)$ (respectively $k_{g,0}(., y)$, $k_{g,1}(., y)$) are orthogonal for the inner product induced by k_b (resp. k_g) but they are not orthogonal for $L^2(D, \mu)$.

2.2. Remarks on the numerical computation of k_0 and k_1

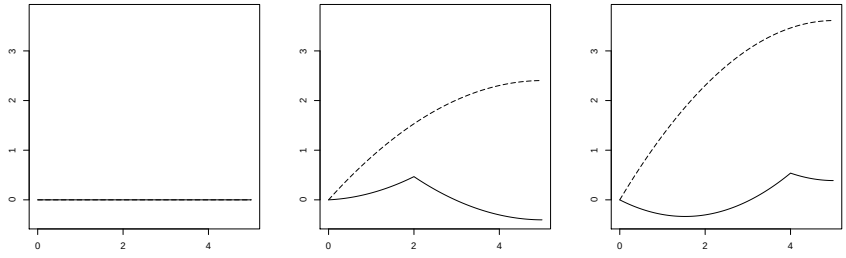
In the previous example, the integrals of Eq. 20 have been approximated using a Riemann sum operator I_R, and we discuss here the influence of this approximation. Let $\hat{k}_0$ be the approximation of k_0 using Riemann sums in Eq. 20. By definition of a RKHS, the evaluation is linear continuous on $\mathcal{H}$ so the operator I_R is also linear continuous and the pattern of the previous proof is still valid when I is replaced by I_R. The function $\hat{k}_0$ corresponds in fact to Proposition 1 applied to the averaging operator I_R. We thus have the following result:

Corollary of Proposition 1: The function $\hat{k}_0$ is the reproducing kernel of the subspace of $\mathcal{H}$ of zero mean functions for I_R.

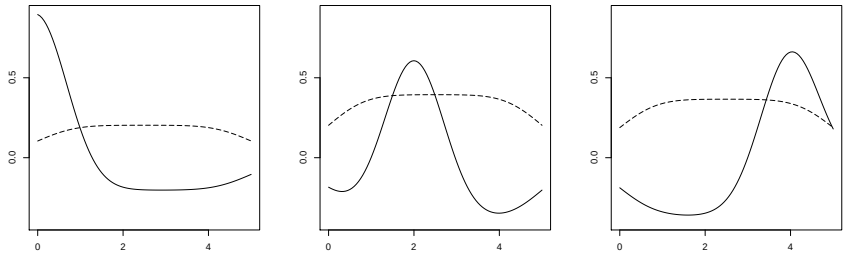
This ensures that the approximation of Eq. 20 by Riemann sums is still a s.p.d. function. For all $x \in [0, 5]$ we thus have $I_R(\hat{k}_0(x, .)) = 0$ and $I(k_0(x, .)) = 0$, but $I_R(k_0(x, .))$ and $I(\hat{k}_0(x, .))$ are generally not equal to zero.

Note that this remark about the Riemann sum is not valid for Monte Carlo approximations of integrals. Indeed, Monte Carlo integration has random output and the use of such methods in Eq. 20 leads to a random function that cannot be interpreted as a kernel (for example the obtained function is not symmetric with probability one).

The kernel k_b can be used to compare the error induced by replacing I by I_R since the integrals of Eq. 20 can be computed analytically. For I_R based



(a) $k_{b,i}(.,0)$ (b) $k_{b,i}(.,2)$ (c) $k_{b,i}(.,4)$



(d) $k_{g,i}(.,0)$ (e) $k_{g,i}(.,2)$ (f) $k_{g,i}(.,4)$

Figure 2: Representation of the kernels $k_{b,i}(.,y)$ and $k_{g,i}(.,y)$ for $y = 0, 2, 4$ and $i = 0, 1$. The dashed lines correspond to $k_{b,1}$, $k_{g,1}$ and the solid lines are for $k_{b,0}$ and $k_{g,0}$. The original kernels $k_b(.,y)$ and $k_g(.,y)$ can be retrieved by summing the two curves.

on a set of 100 points equally spaced over $[0, 5]$, the maximum distance we observed between $k_{b,0}$ and $\hat{k}_{b,0}$ is lower than 0.02 so the two functions would be indistinguishable in Figure 2. Hereafter, the integrals will always be computed using Riemann sum approximations and we will not differentiate k_0 and $\hat{k}_0$ any more.

2.3. Generalization for multi-dimensional RKHS

The former decomposition of one-dimensional kernels leads directly to the decomposition of tensor product kernels if the k^i's satisfy H1:

$$K(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{d} k^i(x_i, y_i) = \prod_{i=1}^{d} (k_0^i(x_i, y_i) + k_1^i(x_i, y_i)). \tag{22}$$

Since $k_0^i(x_i, y_i)$ is a kernel of a one-dimensional RKHS, its effect is similar to the constant function 1 in an ANOVA kernel. This decomposition highlights the similarity between usual tensor product kernels (power-exponential, Brownian, Matérn...) and ANOVA kernels.

3. A new class of kernels for sensitivity analysis

We now propose a special class of ANOVA kernels:

$$K^*_{ANOVA}(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{d}(1+k_0^i(x_i,y_i)) = 1 + \sum_{I\subset\{1,\dots,d\}}\prod_{i\in I} k_0^i(x_i,y_i), \tag{23}$$

where the k_0^i are obtained by decomposing kernels as in the previous section. We now develop the use of such kernels for the ANOVA representation of the best predictor and for the computation of global sensitivity indices.

3.1. ANOVA representation of m

Proposition 2. *If m is a best predictor based on K^*_{ANOVA},*

$$m_I = \left(\bigodot_{i\in I}\mathbf{k}_0^i(x_i)\right)^t \mathrm{K}^{-1}\mathbf{F} \tag{24}$$

is the term of the functional ANOVA representation of m indexed by I. Hence, the decomposition of m given by Eq. 17 coincides with its functional ANOVA representation (Eq. 4).

Proof. The kernels k_0^i are associated with RKHS $\mathcal{H}_0^i$ of zero-mean functions, so we have $\mathbb{1}^i \perp_{L^2} \mathcal{H}_0^i$. The underlying RKHS associated with K^*_{ANOVA} is

$$\mathcal{H}^*_{ANOVA} = \prod_{i=1}^{d}\left(\mathbb{1}^i \overset{\perp}{\oplus} \mathcal{H}_0^i\right) \tag{25}$$

where $\perp$ stands for the L^2 inner product. The result follows. □

Unlike usual ANOVA kernels, the class of K^*_{ANOVA} ensures that the terms m_I are mutually orthogonal in the L^2 sense.

Example 2: ANOVA representation. In order to illustrate the use of the K^*_{ANOVA} kernels, we consider an analytical test function called the g-function of Sobol [14]. This function is defined over $[0,1]^d$ by

$$g(x_1,\dots,x_d) = \prod_{k=1}^{d}\frac{|4x_k-2|+a_k}{1+a_k} \qquad \text{with } a_k>0. \tag{26}$$

We consider here the case $d=2$ and we set $a_1=1,\ a_2=2$. Starting from a one-dimensional Matérn 3/2 kernel

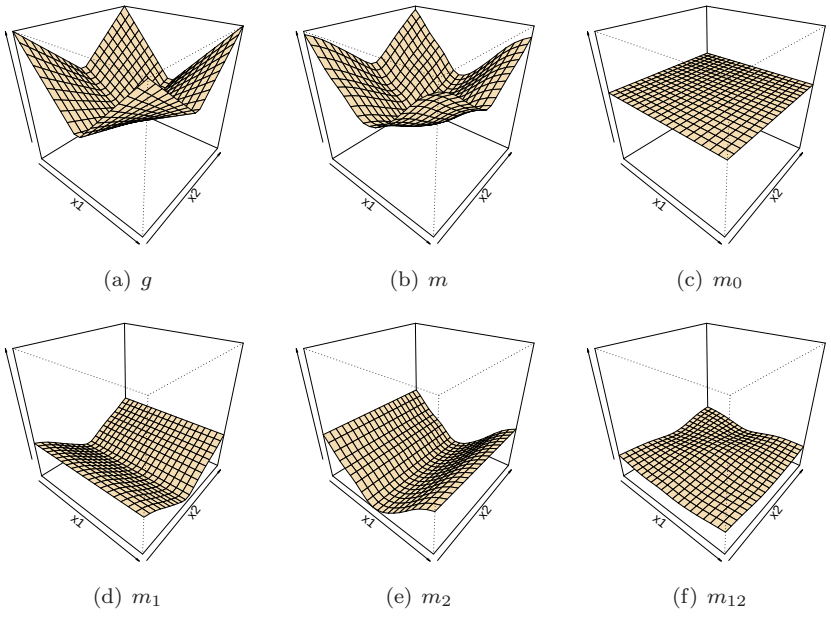
$$k_m(x,y) = (1+2\,|x-y|)\exp(-2\,|x-y|), \tag{27}$$

we derive the expression of K^*_{ANOVA} using Eq. 20 and 23:

$$K^*_{ANOVA}(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{2}\left(1 + k_m(x_i,y_i) - \frac{\int_0^1 k_m(x_i,s)\mathrm{d}\mu(s)\int_0^1 k_m(y_i,s)\mathrm{d}\mu(s)}{\iint_0^1 k_m(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)}\right). \tag{28}$$

We then build the optimal interpolator $m \in \mathcal{H}^*_{ANOVA}$ based on the observation of g at 20 points of $[0,1]^2$ (which result from a LHS-maximin procedure [16]). According to Proposition 2, the function m can be split as a sum of 4 submodels m_0, m_1, m_2 and m_{12} which are represented in Figure 3. We observe numerically that the mean value of m_1, m_2 and m_{12} is lower than 10^{-14} (in absolute value), corroborating that these functions are zero-mean. Similarly, the value of the numerical computation of the scalar products between any two different functions of the set $\{m_0, m_1, m_2, m_{12}\}$ has the same order of magnitude.

As a comparison, we consider on this test function the usual ANOVA kernel $K_{ANOVA}(\mathbf{x},\mathbf{y}) = (1 + k_m(x_1,y_1)) \times (1 + k_m(x_2,y_2))$. For the corresponding model, the mean value of m_1, m_2 and m_{12} is typically in the range $[-0.1, 0.3]$ and a scalar product between m_I and m_J in $[-0.01, 0.03]$ for $I \neq J$.



(a) g (b) m (c) m_0

(d) m_1 (e) m_2 (f) m_{12}

Figure 3: Representation of the g-function, the model m and all the submodels on $[0,1]^2$. The z scale is the same on all graphs.

3.2. Computation of sensitivity indices

We have seen that the terms m_I of the ANOVA representation of m can be obtained naturally when using kernel K^*_{ANOVA}. As the sensitivity indices are based on those terms, they also benefit from the structure of K^*_{ANOVA} and they can be computed analytically and efficiently:

Proposition 3. *The sensitivity indices S_I of m are given by:*

$$S_I = \frac{\mathrm{var}(m_I(\mathbf{X}_I))}{\mathrm{var}(m(\mathbf{X}))} = \frac{\mathbf{F}^t \mathrm{K}^{-1} \left(\bigodot_{i\in I} \Gamma_i\right) \mathrm{K}^{-1}\mathbf{F}}{\mathbf{F}^t \mathrm{K}^{-1} \left(\bigodot_{i=1}^{d} (1_{n\times n} + \Gamma_i) - 1_{n\times n}\right) \mathrm{K}^{-1}\mathbf{F}} \tag{29}$$

where Γ_i is the $n \times n$ matrix $\Gamma_i = \int_{D_i} \mathbf{k}_0^i(x_i)\mathbf{k}_0^i(x_i)^t \mathrm{d}\mu_i(x_i)$ and $1_{k\times l}$ is the $k \times l$ matrix of ones.

Proof. The numerator is obtained by direct calculation. Using that $\mathbf{k}_0^i$ is a vector of zero mean functions and the independence of the X_i we have:

$$\begin{aligned}
\mathrm{var}(m_I(\mathbf{X}_I)) &= \mathrm{var}\left(\left(\bigodot_{i\in I} \mathbf{k}_0^i(X_i)^t\right) \mathrm{K}^{-1}\mathbf{F}\right) \\
&= \mathbf{F}^t \mathrm{K}^{-1} \mathrm{cov}\left(\bigodot_{i\in I} \mathbf{k}_0^i(X_i)\right) \mathrm{K}^{-1}\mathbf{F} \\
&= \mathbf{F}^t \mathrm{K}^{-1} \bigodot_{i\in I} \mathrm{cov}\left(\mathbf{k}_0^i(X_i)\right) \mathrm{K}^{-1}\mathbf{F} \\
&= \mathbf{F}^t \mathrm{K}^{-1} \bigodot_{i\in I} \left(\int_{D_i} \mathbf{k}_0^i(x_i)\mathbf{k}_0^i(x_i)^t \mathrm{d}\mu_i(x_i)\right) \mathrm{K}^{-1}\mathbf{F}.
\end{aligned} \tag{30}$$

For the denominator, we obtain similarly

$$\begin{aligned}
\mathrm{var}(m(\mathbf{X})) = \mathbf{F}^t \mathrm{K}^{-1} \bigodot_{i\in I} &\left(\int_{D_i} \left(1_{n\times 1} + \mathbf{k}_0^i(x_i)\right)\left(1_{n\times 1} + \mathbf{k}_0^i(x_i)\right)^t \mathrm{d}\mu_i(x_i)\right) \mathrm{K}^{-1}\mathbf{F} \\
&- \mathbf{F}^t \mathrm{K}^{-1} 1_{n\times n} \mathrm{K}^{-1}\mathbf{F}.
\end{aligned} \tag{31}$$

We then use the property that the $k_0^i(x,.)$ are zero mean functions again to obtain $\int_{D_i} \left(1_{n\times 1} + \mathbf{k}_0^i(x_i)\right)\left(1_{n\times 1} + \mathbf{k}_0^i(x_i)\right)^t \mathrm{d}\mu_i(x_i) = 1_{n\times n} + \Gamma_i$. □

A similar property is given in [5] for usual separable kernels but the computation of S_I require computation of all S_J for $J \subset I$. This recursivity induces a propagation of the numerical error when the integrals in Γ_i are approximated numerically. Here, the particular structure of K^*_{ANOVA} avoids any recursivity and each index is given directly by Proposition 3.

Example 3: Computation of sensitivity indices. As previously, we consider the g-function as a test function to illustrate Proposition 3. One particular asset of g is that the sensitivity indices S_I can be obtained analytically:

$$S_I = \frac{\prod_{i \in I} \frac{1}{3(1+a_i)^2}}{\prod_{k=1}^{d} \left(1 + \frac{1}{3(1+a_k)^2}\right) - 1}. \tag{32}$$

The value of the dimension is set to $d = 5$ and the coefficients are arbitrarily chosen to be $(a_1, a_2, a_3, a_4, a_5) = (0.2, 0.6, 0.8, 100, 100)$. With these values, the variables X_1, X_2 and X_3 explain 99.99% of the variance of $g(\mathbf{X})$.

Now, we compare the indices given by Proposition 3 for three different kernels versus the analytical value given by Eq. 32. The kernels we use are the Matérn 3/2 kernel k_m introduced in Eq. 27, the Gaussian kernel k_g of Eq. 12 and the kernel $\tilde{k}_b(x, y) = 1 + \min(x, y)$. In the latter, the value 1 is added to the kernel of the Brownian motion to relax the constraint of nullity at the origin.

The Design of Experiment (DoE) is a 50 point maximin LHS over $[0, 1]^5$. As each DoE leads to a particular model and a different value of the sensitivity indices, computation of the indices was repeated for 50 different DoEs. The obtained mean values and the standard deviations are gathered in Table 1. As for example 1, we studied for the model based on $\tilde{k}_b$ the error due to the approximation of the integrals in Eq. 29 by Riemann sums based on 100 points per dimension. The error on the final values of the indices does not appear to be significant since it is typically around 10^{-5}.

I	1	2	3	{1,2}	{1,3}	{2,3}	{1,2,3}
Analytical	0.43	0.24	0.19	0.06	0.04	0.03	0.01
$\tilde{k}_b$	0.44 (0.05)	0.27 (0.05)	0.20 (0.04)	0.01 (0.01)	0.01 (0.00)	0.01 (0.01)	0.00 (0.00)
k_m	0.44 (0.06)	0.24 (0.05)	0.19 (0.04)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.00 (0.00)
k_g	0.33 (0.08)	0.19 (0.06)	0.14 (0.05)	0.01 (0.02)	0.02 (0.02)	0.03 (0.02)	0.03 (0.02)

Table 1: Mean value and standard deviation of the sensitivity indices for the g-function.

The computational accuracy of the global sensitivity indices is judged to be satisfactory for $\tilde{k}_b$ and k_m. On the other hand, the model based on k_g performs significantly worse than the others. This can be explained by the unsuitability of the Gaussian kernel for the approximation of a function that is not differentiable everywhere. This example highlights the importance of choosing a kernel suited to the problem at hand.

Although the indices are correctly estimated with $\tilde{k}_b$ and k_m, the sum of the indices presented in Table 1 is respectively equal to 0.94 and 0.90 for these kernels (whereas we obtain 1 for the first line). It implies that the associated models

present variations for some sets of indices that are not represented in this table and eventually the sparsity in the variance structure of the g-function is not perfectly captured by the model. As the kernels K^*_{ANOVA} are ANOVA kernels, methods such as SUPANOVA [9] or hierarchical kernel learning [3] can be applied for detecting this sparsity. The combination of K^*_{ANOVA} and SUPANOVA present similarities with the COSSO [11] for splines but such developments are out of the scope of this article.

4. Generalization to regularization problems

We have considered for now the interpolation point of view and the expression of the best predictor given in Eq. 9. However, all the results can immediately be generalized to regularization problems. Indeed, the proofs of Propositions 2 and 3 are still valid if we replace the Gram matrix K by $\mathrm{K} + \lambda \mathrm{I}$ so we have the following results.

Corollary of Proposition 2: *The terms of the ANOVA representation of the regularization best predictor* $\tilde{m}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^t(\mathrm{K} + \lambda \mathrm{I})^{-1}\mathbf{F}$ *are*

$$\tilde{m}_I = \left(\bigodot_{i \in I} \mathbf{k}_0^i(x_i)^t \right) (\mathrm{K} + \lambda \mathrm{I})^{-1}\mathbf{F}. \tag{33}$$

Corollary of Proposition 3: *The expression of the sensitivity indices of* $\tilde{m}$ *is*

$$S_I = \frac{\mathbf{F}^t(\mathrm{K} + \lambda \mathrm{I})^{-1} \left(\bigodot_{i \in I} \Gamma_i \right) (\mathrm{K} + \lambda \mathrm{I})^{-1}\mathbf{F}}{\mathbf{F}^t(\mathrm{K} + \lambda \mathrm{I})^{-1} \left(\bigodot_{i=1}^{d} (1_{n \times n} + \Gamma_i) - 1_{n \times n} \right) (\mathrm{K} + \lambda \mathrm{I})^{-1}\mathbf{F}}. \tag{34}$$

Example 4. The aim of this last example is to illustrate two points: the corollaries we have just stated and the use of a non-uniform measure. We consider here a simple test function introduced in [15]:

$$\begin{aligned} f \ &: \mathbb{R}^2 \to \mathbb{R} \\ &\mathbf{x} \mapsto x_1 + x_2^2 + x_1 x_2. \end{aligned} \tag{35}$$

The input space $\mathbb{R} \times \mathbb{R}$ is endowed with a measure $\mu_1 \otimes \mu_2$ where μ_i is the standard normal distribution. With these settings, the ANOVA representation of f is

$$f(\mathbf{x}) = \underbrace{1}_{f_0} + \underbrace{x_1}_{f_1} + \underbrace{x_2^2 - 1}_{f_2} + \underbrace{x_1 x_2}_{f_{1,2}} \tag{36}$$

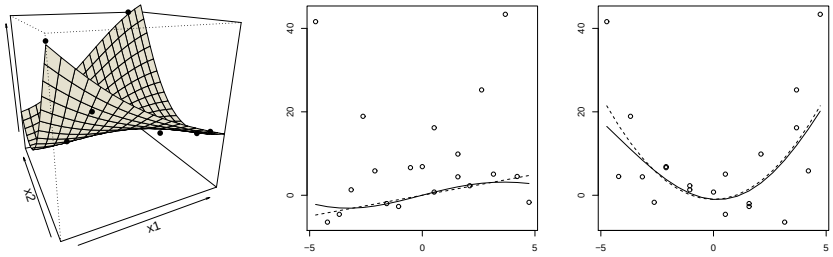
and the sensitivity indices S_1, S_2 and $S_{1,2}$ are respectively equal to 1/4, 1/2, 1/4. As we want to illustrate here a regularization issue, we assume that the observations of f are affected by a noise of observation $f_{obs} = f + \varepsilon$ where ε is a specific realization of a Gaussian white noise with variance λ.

The DoE is a 20 point LHS maximin over $[-5,5]^2$ and we consider the following kernel over $\mathbb{R}^2 \times \mathbb{R}^2$:

$$K^*_{ANOVA}(\mathbf{x},\mathbf{y}) = 200\,(1+g_0(x_1,y_1))\,(1+g_0(x_2,y_2)) \tag{37}$$

where g_0 is obtained by decomposing a Gaussian kernel $k_g(x,y) = exp\left(-\left(\frac{x-y}{10}\right)^2\right)$. The value 200 is a scale parameter roughly estimated from the variance of the observations at the DoE.

Figure 4 represents the model and main effects we obtained for $\lambda = 1$. It appears that despite the small number of design points, the submodels $\tilde{m}_1$ and $\tilde{m}_2$ give a good approximation of the analytical main effects f_1 and f_2.



(a) $\tilde{m}(\mathbf{x})$ (b) $\tilde{m}_1(x_1)$ (c) $\tilde{m}_2(x_2)$

Figure 4: Representations of the model and two submodels for $\lambda = 1$. The dots correspond to the observations and for panels (b) and (c), the dashed lines stand for the analytical main effects given in Eq. 36.

Table 2 gives the values of the sensitivity indices for different levels of noise. As $\mathrm{var}(f(\mathbf{X})) = 4$ and $\mathrm{var}(\varepsilon(\mathbf{X})) = \lambda$, the variance of the noise we introduced is up to four times larger than the variance of the signal we want to extract. The computation of the indices appears to be very robust for this example, even for a low signal-to-noise ratio.

For $\lambda = 0$, the model retrieves the analytical value of the parameters. When λ increases, the standard deviation of the estimates increases and the index S_2 tends to be slightly underestimated. One can note that the analytical value is at the most one standard deviation away from the estimated mean.

5. Concluding remarks

We introduced in this article a special case of ANOVA kernels with particular univariate kernels so that the L^2-orthogonality to constants is respected. This new class of kernels offers good interpretation properties to the models since their functional ANOVA decomposition can be obtained analytically, without the usual recursive integral calculations. Finally, we showed that this natural decomposition of the best predictor benefits the computation of the global sensitivity indices.

I	1	2	{1,2}
Analytical	0.25	0.5	0.25
$\lambda = 0$	0.25 (0.00)	0.5 (0.00)	0.25 (0.00)
$\lambda = 1$	0.25 (0.05)	0.48 (0.04)	0.28 (0.03)
$\lambda = 2$	0.26 (0.06)	0.47 (0.05)	0.28 (0.04)
$\lambda = 4$	0.24 (0.08)	0.46 (0.06)	0.30 (0.05)
$\lambda = 8$	0.26 (0.10)	0.44 (0.08)	0.30 (0.07)
$\lambda = 16$	0.28 (0.14)	0.40 (0.11)	0.32 (0.09)

Table 2: Mean value and standard deviation of the sensitivity indices for the test function f for various values of the noise parameter λ.

The use of Riemann sums for the numerical computation of the integrals has been discussed and the error of approximation has shown to be negligible in various examples. Up to calculation or tabulation of the integral of univariate kernels, replacing the usual ANOVA kernels with the ones proposed here may be done at negligible cost in applications, with substantial benefits for the model interpretability and global sensitivity analysis studies.

The parameter estimation issue for K^*_{ANOVA} has not been raised yet in this article. This is however an important point for the practical use of those kernels. The use of the likelihood theory has been considered, but many points such as the links between the optimal parameters for K and the optimal parameters for the associated K^*_{ANOVA} need to be studied in detail. Finally, since the pattern of the proof of Proposition 1 can be applied to any bounded operator on $\mathcal{H}$, lines of future research include focusing on other operators than the integral operator I, for example for building RKHS respecting orthogonality to a family of trend basis functions.

Acknowledgments

The author would like to thank the reviewers for their valuable comments. Their constructive remarks contributed significantly to the actual version of the article. We are also grateful to James Hensman and Ben Youngman for all their remarks and comments.

[1] A. Antoniadis. Analysis of variance on function spaces. Statistics, 15:59–71, 1984.

[2] N. Aronszajn. Theory of reproducing kernels. Transaction of the American Mathematical Society, 68, Issue 3:337 – 404, 1950.

[3] F. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical report, INRIA - WILLOW Project-Team. Laboratoire d'Informatique de l'Ecole Normale Supérieure, 2009.

[4] Alain Berlinet and Christine Thomas-Agnan. Reproducing kernel Hilbert spaces in probability and statistics. Kluwer Academic Publishers, 2004.

[5] W. Chen, R. Jin, and A. Sudjianto. Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. Journal of mechanical design, 127, 2005.

[6] David Duvenaud, Hannes Nickisch, and Carl Edward Rasmussen. Additive Gaussian processes. In Advances in Neural Information Processing Systems 25, pages 1–8, Granada, Spain, December 2011.

[7] B. Efron and C. Stein. The jackknife estimate of variance. The Annals of Statistics, 9:586–596, 1981.

[8] C. Gu. Smoothing spline ANOVA models. Springer Verlag, 2002.

[9] S.R. Gunn and J.S. Kandola. Structural modelling with sparse kernels. Machine learning, 48:137–163, 2002.

[10] P. Krée. Produits tensoriels complétés d'espaces de Hilbert. Séminaire Paul Krée, Vol 1:No. 7, 1974–1975.

[11] Y. Lin and H.H. Zhang. Component selection and smoothing in multivariate nonparametric regression. The Annals of Statistics, 34:2272–2297, 2006.

[12] A. Marrel, B. Iooss, B. Laurent, and O. Roustant. Calculations of sobol indices for the gaussian process metamodel. Reliability Engineering & System Safety, 94:742–751, 2009.

[13] C.E. Rasmussen and C. Williams. Gaussian Processes for Machine Learning. MIT Press, 2006.

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

[15] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. Global sensitivity analysis: the primer. Wiley Online Library, 2008.

[16] T.J. Santner, B.J. Williams, and W. Notz. The design and analysis of computer experiments. Springer Series in Statistics, 2003.

[17] I.M. Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and Computers in Simulation, 55:271–280, 2001.

[18] Mark O. Stitson, Alex Gammerman, Vladimir Vapnik, Volodya Vovk, Chris Watkins, and Jason Weston. Support vector regression with ANOVA decomposition kernels. Technical report, Royal Holloway , University of London, 1997.

[19] B.K.S. Vangeepuram. Reproducing Kernel Space Embeddings and Metrics on Probability Measures. PhD thesis, University of California, San Diego, 2010.

[20] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. The Annals of Statistics, 23:1865–1895, 1995.