Some kernels for computer experiments

Olivier Roustant

École Nationale Supérieure des Mines de Saint-Étienne

2011/12 – ERCIM
Outline

1 Methods: ANOVA kernels for the global sensitivity analysis of Kriging metamodels
   - The aim: To perform the global SA without the curse of recursion
   - Global SA of "1+0" separable functions
   - From "1+0" separable functions to ANOVA kernels via RKHS
   - The probabilistic interpretation

2 Software: New kernel classes in R packages
   - Block-additive kernels (R package "fanovaGraph")
   - Non-stationary kernels based on scaling transform [DiceKriging]

3 References
### Associated collaborators

- Joint work with N. Durrande, D. Ginsbourger and L. Carraro

### Methods

- Joint work with N. Durrande, D. Ginsbourger and L. Carraro

### Software

- Block-additive kernels: Joint works with J. Fruth, T. Muehlenstaedt, S. Kuhnt and L. Carraro.
- Scaling transform: Joint work with D. Ginsbourger and Y. Deville.
FANOVA decomposition and recursion issue

**Property and definition (see Hoeffding [7] or Efron and Stein [5])**

Let $x_1, ..., x_d$ be independent random variables, with distribution $d\nu(x) = d\nu_1(x_1) \ldots d\nu_d(x_d)$ and $f$ an integrable function. Then $f$ admits a unique "FANOVA" decomposition:

$$f(x) = \mu_0 + \sum_{j=1}^{d} \mu_j(x_j) + \sum_{j<k} \mu_{j,k}(x_j, x_k) + \sum_{j<k<l} \mu_{j,k,l}(x_j, x_k, x_l) + \cdots + \mu_{1,\ldots,d}(x_1, \ldots, x_d)$$

where each term is centered and satisfies $E(\mu_I(x_I)|x_J) = 0$ for all $J \subset I$. 

FANOVA decomposition and recursion issue

Property and definition (see Hoeffding [7] or Efron and Stein [5])

Let $x_1, \ldots, x_d$ be independent random variables, with distribution $d\nu(x) = d\nu_1(x_1) \ldots d\nu_d(x_d)$ and $f$ an integrable function. Then $f$ admits a unique "FANOVA" decomposition:

$$f(x) = \mu_0 + \sum_{j=1}^{d} \mu_j(x_j) + \sum_{j<k} \mu_{j,k}(x_j, x_k)$$

$$+ \sum_{j<k<l} \mu_{j,k,l}(x_j, x_k, x_l) + \cdots + \mu_{1,\ldots,d}(x_1, \ldots, x_d)$$

where each term is centered and satisfies $E(\mu_I(x_l) | x_J) = 0$ for all $J \subsetneq I$.

Orthogonality

In particular the $\mu_I$'s are orthogonal.
Context: FANOVA decomposition and recursion issue

Recursion issue

We have:

- \( \mu_0 = E[f(x)] \)
- \( \mu_k(x_k) = E[f(x)|x_k] - \mu_0 \)
- \( \mu_{j,k}(x_j, x_k) = E[f(x)|x_j, x_k] - \mu_j(x_j) - \mu_k(x_k) - \mu_0 \)
- \( \ldots \)

In general, the terms are expressed recursively by:

\[
\mu_J(x_J) = E[f(x)|x_J] - \sum_{J' \subsetneq J} \mu_{J'}(x_{J'})
\]
Context: FANOVA decomposition and recursion issue

Recursion issue

We have:

\[ \mu_0 = E[f(x)] \]
\[ \mu_k(x_k) = E[f(x)|x_k] - \mu_0 \]
\[ \mu_{j,k}(x_{j,k}) = E[f(x)|x_j, x_k] - \mu_j(x_j) - \mu_k(x_k) - \mu_0 \]
\[ \ldots \]

In general, the terms are expressed **recursively** by:

\[ \mu_J(x_J) = E[f(x)|x_J] - \sum_{J' \subsetneq J} \mu_{J'}(x_{J'}) \]

Variance decomposition and Sobol indices

Orthogonality implies: \( V = \sum V_i \), with \( V_i = var(\mu_i(x_i)) \)

The **Sobol indices** are defined as variance ratios \( S_i := V_i/V \)
The aim

Global SA of the Kriging mean

- When the kernel is separable, i.e. \( k(x, y) = \prod_{i=1}^{d} k^i(x_i, y_i) \), the global SA of the Kriging mean is performed analytically ([9], [2])
- However, as for the definition, we face the curse of recursion
The aim

Global SA of the Kriging mean
- When the kernel is separable, i.e. $k(x, y) = \prod_{i=1}^{d} k^i(x_i, y_i)$, the global SA of the Kriging mean is performed analytically ([9], [2])
- However, as for the definition, we face the curse of recursion

The problem
- To find a kernel s.t. the SA of the Krig. mean is done without recursion
A favorable case: "1+0" separable functions

Proposition (at least: Sobol, 2003 [10])

Let $\nu = \nu_1 \ldots \nu_d$ a measure on $\mathbb{R}^d$, and $f$ be a separable function:

$$f(x) = \prod_{i=1}^{d} (1 + f_i(x_i))$$

such that the $f_i$’s are zero-mean functions: $\int f_i(x_i) d\nu_i(x_i) = 0$. Then:
A favorable case: "1+0" separable functions

Proposition (at least: Sobol, 2003 [10])

Let $\nu = \nu_1 \ldots \nu_d$ a measure on $R^d$, and $f$ be a separable function:

$$f(x) = \prod_{i=1}^{d} (1 + f_i(x_i))$$

such that the $f_i$'s are zero-mean functions: $\int f_i(x_i) d\nu_i(x_i) = 0$. Then:

- The FANOVA decomposition is obtained - without any recursion - by expanding the product: $\mu_I(x_I) = \prod_{i \in I} f_i(x_i)$
A favorable case: "1+0" separable functions

Proposition (at least: Sobol, 2003 [10])

Let $\nu = \nu_1 \ldots \nu_d$ a measure on $\mathbb{R}^d$, and $f$ be a separable function:

$$f(x) = \prod_{i=1}^{d} (1 + f_i(x_i))$$

such that the $f_i$’s are zero-mean functions: $\int f_i(x_i) d\nu_i(x_i) = 0$. Then:

- The FANOVA decomposition is obtained - without any recursion - by expanding the product: $\mu_I(x_I) = \prod_{i \in I} f_i(x_i)$
- The Sobol indices are given analytically by $V_I/V$, where:

$$V_I = \prod_{i \in I} \int f_i^2(x_i) d\nu_i(x_i) \quad \text{and} \quad V = \prod_{i=1}^{d} \left( 1 + \int f_i^2(x_i) d\nu_i(x_i) \right) - 1$$
Sensitivity analysis of the Kriging mean

Consider the Kriging mean $m(x) = k(x)^t K^{-1} F$, with:

- $k(x) := (k(x, x^{(j)}))_{1 \leq j \leq n}$: cov. between $x$ and design points
- $K := (k(x^{(j)}, x^{(j')}))_{1 \leq j, j' \leq n}$: covariances between design points

Then, with $\alpha = K^{-1} F$, we have: $m(.) = \sum_{j=1}^{n} \alpha_j k(., x^{(j)})$
Sensitivity analysis of the Kriging mean

Consider the Kriging mean \( m(x) = k(x)^t K^{-1} F \), with:
- \( k(x) := (k(x, x^{(j)}))_{1 \leq j \leq n} \) : covariance between \( x \) and design points
- \( K := (k(x^{(j)}, x^{(j')})_{1 \leq j, j' \leq n} \) : covariances between design points

Then, with \( \alpha = K^{-1} F \), we have:
\[
m(.) = \sum_{j=1}^{n} \alpha_j k(., x^{(j)})
\]

Sufficient condition (SC) to avoid recursion

To find \( k \) s.t. all the \( k(., y) \) are "1+0" separable functions
From functions to kernels

RKHS - Main facts [case of real-valued functions]

- A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is a Hilbert space of functions s.t. the evaluations $f \rightarrow f(x)$ are continuous for all $x$. 

By Riesz theorem, there exists $k(x, \cdot) \in \mathcal{H}$ s.t. $f(x) = \langle f, k(x, \cdot) \rangle$.

In particular, with $f = k(y, \cdot)$:

$$k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle$$

$k$ is a kernel:

$$\sum a_i a_j k(x_i, x_j) = \sum a_i a_j \langle k(x_i, \cdot), k(x_j, \cdot) \rangle \geq 0$$

Equivalence between kernels and RKHS: the Moore-Aronszajn theorem

For any kernel $k$, there exists a unique RKHS with kernel $k$: $\mathcal{H}_k := \text{span}(k(x, \cdot), x)$, with $\langle k(x, \cdot), k(y, \cdot) \rangle = k(x, y)$.
From functions to kernels

RKHS - Main facts [case of real-valued functions]

- A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is a Hilbert space of functions s.t. the evaluations $f \mapsto f(x)$ are continuous for all $x$.
- By Riesz theorem, $\exists! k(., x) \in \mathcal{H}$ s.t. $f(x) = \langle f, k(x, .) \rangle$
From functions to kernels

RKHS - Main facts [case of real-valued functions]

- A reproducing kernel Hilbert space (RKHS) \( \mathcal{H} \) is a Hilbert space of functions s.t. the evaluations \( f \rightarrow f(x) \) are continuous for all \( x \).
- By Riesz theorem, \( \exists ! k(., x) \in \mathcal{H} \) s.t. \( f(x) = \langle f, k(x, .) \rangle \)
- In particular, with \( f = k(y, .) \) : \( k(x, y) = \langle k(x, .), k(y, .) \rangle \)
From functions to kernels

RKHS - Main facts [case of real-valued functions]

- A reproducing kernel Hilbert space (RKHS) \( \mathcal{H} \) is a Hilbert space of functions s.t. the evaluations \( f \rightarrow f(x) \) are continuous for all \( x \).
- By Riesz theorem, \( \exists! k(., x) \in \mathcal{H} \) s.t. \( f(x) = \langle f, k(x, .) \rangle \)
- In particular, with \( f = k(y, .) \) : \( k(x, y) = \langle k(x, .), k(y, .) \rangle \)
- \( k \) is a kernel:
  \[
  \sum a_ia_jk(x_i, x_j) = \sum a_ia_j\langle k(x_i, .), k(x_j, .) \rangle = \| \sum a_i k(x_i, .) \|^2 \geq 0
  \]
From functions to kernels

**RKHS - Main facts [case of real-valued functions]**

- A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is a Hilbert space of functions s.t. the evaluations $f \mapsto f(x)$ are continuous for all $x$.
- By Riesz theorem, $\exists! k(., x) \in \mathcal{H}$ s.t. $f(x) = \langle f, k(x, .) \rangle$
- In particular, with $f = k(y, .)$: $k(x, y) = \langle k(x, .), k(y, .) \rangle$
- $k$ is a kernel:
  $$\sum a_ia_jk(x_i, x_j) = \sum a_ia_j\langle k(x_i, .), k(x_j, .) \rangle = \| \sum a_i k(x_i, .) \|^2 \geq 0$$

**Equivalence between kernels and RKHS : the Moore-Aronszajn theorem**

For any kernel $k$, there exists a unique RKHS with kernel $k$:
$$\mathcal{H}_k := \text{span}(k(x, .), x), \text{ with } \langle k(x, .), k(y, .) \rangle = k(x, y)$$
RKHS of zero-mean functions

Proposition

Let $\mathcal{H}$ be a RKHS of kernel $k$, satisfying $\int \sqrt{k(s, s)} d\nu(s) < \infty$. Then:

- $\exists! R$ s.t. $\int h(x) d\nu(x) = \langle R, h \rangle_\mathcal{H}$.
RKHS of zero-mean functions

**Proposition**

Let $\mathcal{H}$ be a RKHS of kernel $k$, satisfying $\int \sqrt{k(s, s)} d\nu(s) < \infty$. Then:

- $\exists! R$ s.t. $\int h(x) d\nu(x) = \langle R, h \rangle_{\mathcal{H}}$.

Denote $\mathcal{H}_1 := \text{span}(R)$ and $\mathcal{H}_0 := \mathcal{H}_1^\perp$. Then:
RKHS of zero-mean functions

Proposition

Let $\mathcal{H}$ be a RKHS of kernel $k$, satisfying $\int \sqrt{k(s, s)}d\nu(s) < \infty$. Then:

- $\exists! R$ s.t. $\int h(x)d\nu(x) = \langle R, h \rangle_{\mathcal{H}}$.

Denote $\mathcal{H}_1 := \text{span}(R)$ and $\mathcal{H}_0 := \mathcal{H}^\perp$. Then:

- $\mathcal{H}_0$ is the subspace of zero-mean functions:
  
  $$h \in \mathcal{H}_0 \iff \langle R, h \rangle_{\mathcal{H}} = 0 \iff \int h(x)d\nu(x) = 0$$

Proof: use the properties of the sum and projection of RKHS ([1]).
RKHS of zero-mean functions

Proposition

Let $\mathcal{H}$ be a RKHS of kernel $k$, satisfying $\int \sqrt{k(s, s)} d\nu(s) < \infty$. Then:

- $\exists! R$ s.t. $\int h(x) d\nu(x) = \langle R, h \rangle_{\mathcal{H}}$.

Denote $\mathcal{H}_1 := \text{span}(R)$ and $\mathcal{H}_0 := \mathcal{H} \perp_1$. Then:

- $\mathcal{H}_0$ is the subspace of zero-mean functions:
  $$h \in \mathcal{H}_0 \iff \langle R, h \rangle_{\mathcal{H}} = 0 \iff \int h(x) d\nu(x) = 0$$

- $\mathcal{H}_1$ and $\mathcal{H}_0$ are RKHS with kernels:
  $$k_1(x, y) = \frac{\int k(x, s) d\nu(s) \int k(y, t) d\nu(t)}{\int \int k(s, t) d\nu(s) d\nu(t)}$$
  $$k_0 = k - k_1$$

Proof: use the properties of the sum and projection of RKHS ([1])
A new class of kernels for sensitivity analysis

Start with an ANOVA kernel on $\mathbb{R}^d$:

$$k_{\text{ANOVA}}(x, y) = \prod_{i=1}^{d} (1 + k^i(x_i, y_i))$$

where the $k^i$ are 1-dimensional kernels. Now, replace each $k^i$ by

$$k_0^i(x, y) = k^i(x, y) - \frac{\int k^i(x, s) d\nu(s) \int k^i(y, t) d\nu(t)}{\int \int k^i(s, t) d\nu(s) d\nu(t)}$$

and obtain:

$$k^*_{\text{ANOVA}}(x, y) = \prod_{i=1}^{d} (1 + k_0^i(x_i, y_i))$$
A new class of kernels for sensitivity analysis

Start with an ANOVA kernel on $\mathbb{R}^d$:

$$k_{\text{ANOVA}}(x, y) = \prod_{i=1}^{d} (1 + k^i(x_i, y_i))$$

where the $k^i$ are 1-dimensional kernels. Now, replace each $k^i$ by

$$k^i_0(x, y) = k^i(x, y) - \frac{\int k^i(x, s)d\nu(s)\int k^i(y, t)d\nu(t)}{\int\int k^i(s, t)d\nu(s)d\nu(t)}$$

$$k^*_{\text{ANOVA}}(x, y) = \prod_{i=1}^{d} (1 + k^i_0(x_i, y_i))$$

Important remark

All the $k^*_{\text{ANOVA}}(\cdot, y)$ are "1+0" separable functions
Denote:
- $m_I$'s the terms of the FANOVA decomposition of $m$
- $S_I$ the corresponding Sobol indices

Then, if $k = k_{\text{ANOVA}}^*$, we have – without any recursion:

- $m_I = \prod_{i \in I} k_0^i (x_i)^t K^{-1} F$
- $S_I = \frac{F^t K^{-1} (\bigodot_{i \in I} \Gamma_i) K^{-1} F}{F^t K^{-1} (\bigodot_{i \in I} (1_{n \times n} + \Gamma_i) - 1_{n \times n}) K^{-1} F}$

with $\Gamma_i = \int k_0^i(x_i) k_0^i(x_i)^t d\nu_i(x_i)$
Illustration with a 2 dimensional g-Sobol function

Estimated Sobol indices: $S_1 = 0.675$, $S_2 = 0.30$, $S_{12} = 0.025$
(rounded true values: 0.69, 0.30, 0.02)
Loève representation theorem [1]

- Let $\mathcal{H}_k$ be the RKHS of kernel $k$
- And $Z$ be a centered Gaussian process with (covariance) kernel $k$.

Then, introducing $\mathcal{L}(Z) = \text{span}(Z_x, x)$, with $\langle Z_x, Z_y \rangle = E(Z_x Z_y)$, the following map defines an isometry:

$$
\phi : \mathcal{H}_k \rightarrow \mathcal{L}(Z)
$$

$$
k(x,) \rightarrow Z_x
$$
Loève representation theorem [1]

- Let $\mathcal{H}_k$ be the RKHS of kernel $k$.
- And $Z$ be a centered Gaussian process with (covariance) kernel $k$.

Then, introducing $\overline{\mathcal{L}}(Z) = \text{span}(Z_x, x)$, with $\langle Z_x, Z_y \rangle = E(Z_x Z_y)$, the following map defines an isometry:

$$\phi : \mathcal{H}_k \to \overline{\mathcal{L}}(Z)$$

$$k(x, .) \to Z_x$$

Application: Translation of $\mathcal{H}_1$ and $\mathcal{H}_0$ (1-d case)

$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_0$ is translated "pointwise" as:

$$Z(.) = E \left( Z(.) \left| \int Z(t) d\nu(t) \right. \right) + Z(.) - E \left( Z(.) \left| \int Z(t) d\nu(t) \right. \right)$$

The kernels of the two corresponding processes are $k_1$ and $k_0$. 
References and connected works

- Wahba [12] has worked on RKHS of zero-mean functions, in the context of smoothing splines.
- Further developments of this work can be found in our paper [4] as well as in the PhD thesis of N. Durrande [3]
Part II. Software : New kernel classes in R packages
Block-additive graph-based kernels – Illustration of the idea

A toy [and advertising] example: Ishigami function

\[ f(x) = \sin(x_1) + A\sin^2(x_2) + Bx_3^4\sin(x_1) = f_2(x_2) + f_{1,3}(x_1, x_3) \]
Block-additive graph-based kernels with \textit{fanovaGraph}

What is implemented?

Scope: functions of high complexity (high order interactions)
- Graph estimation (from data)
- Graph thresholding
- Kernel construction (from the graph)
- Kriging model inference with a block-additive kernel
- Prediction

To go further and see more complex case studies...
- Attend the talk of J. Fruth at ERCIM!
Scaling-based kernels – illustration

Left: Kriging with a stationary kernel. Right: scaling-based kernel.
Scaling-based kernels

When non-stationarity comes from the input space

Assume that $Y = Z_{g(x)}$ with:

- $Z$ a (stationary) process on $X$, with kernel $K_{\Theta}$
- $g$ a function on $X$. 

The idea is to estimate $g$ (and the parameters $\Theta$) from the data.

Parametrization of $g$ (Xiong, Chen, Apley and Ding, 2007 [13])

$g$: Coordinatewise the antiderivative of a piecewise affine function
Scaling-based kernels

When non-stationarity comes from the input space

Assume that $Y = Z_g(x)$ with:

- $Z$ a (stationary) process on $X$, with kernel $K_\Theta$
- $g$ a function on $X$.

Then the kernel of $Y$ is:

$$K_g(x, y) = \text{cov}(Z_g(x), Z_g(y)) = K(g(x), g(y))$$

The idea is to estimate $g$ (and the parameters $\Theta$) from the data.
Scaling-based kernels

When non-stationarity comes from the input space

Assume that \( Y = Z_g(x) \) with:

- \( Z \) a (stationary) process on \( X \), with kernel \( K_\Theta \)
- \( g \) a function on \( X \).

Then the kernel of \( Y \) is:

\[
K_g(x, y) = \text{cov}(Z_g(x), Z_g(y)) = K(g(x), g(y))
\]

The idea is to estimate \( g \) (and the parameters \( \Theta \)) from the data.

Parametrization of \( g \) (Xiong, Chen, Apley and Ding, 2007 [13])

\( g \): Coordinatewise the antiderivative of a piecewise affine function
Scaling transform with function \textit{km}

**What is implemented?**

- \textit{knots} : possibly a different number of knots per dimension, fixed by the user. Default is two knots (0 and 1) per dimension.

- \textit{values at knots} : estimated by ML (default), or fixed

\textit{Numerical estimation is performed with analytical gradient (BFGS or genetic algorithm genoud)}
Scaling transform with function *km*

**What is implemented?**

- **knots**: possibly a different number of knots per dimension, fixed by the user. Default is two knots (0 and 1) per dimension.
- **values at knots**: estimated by ML (default), or fixed. *Numerical estimation is performed with analytical gradient (BFGS or genetic algorithm genoud)*

**Argument details**

- **scaling**: to be turned to TRUE for scaling transform.
- **knots**: an optional list containing the vectors of knots per dim.
- **coef.cov**: an optional list containing the vectors of values at knots when fixed. [Then *coef.trend* and *coef.var* must be specified]
A. Berlinet and C. Thomas-Agnan.  
*Reproducing kernel Hilbert spaces in probability and statistics.*  

W. Chen, R. Jin, and A. Sudjianto.  
Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty.  

N. Durrande.  

Reproducing kernels for spaces of zero mean functions.  
application to sensitivity analysis.  
*Submitted*, +2011.

B. Efron and C. Stein.  
The jackknife estimate of variance.  
Total interaction indices for the decomposition of functions with high complexity. 
*Submitted, +2011.*

W.F. Hoeffding.
A class of statistics with asymptotically normal distributions. 

Data-driven kriging models based on fanova-decomposition. 

T.J. Santner, B. Williams, and W. Notz.
*The Design and Analysis of Computer Experiments.*

I. Sobol.
Theorems and examples on high dimensional model representation.
S. Touzani.
*Response surface methods based on analysis of variance expansion for sensitivity analysis.*

G. Wahba.
*Spline Models for Observational Data.*

A non-stationary covariance-based kriging method for metamodelling in engineering design.