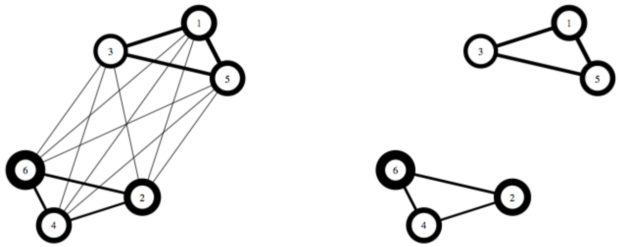
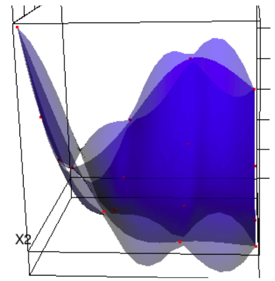




Block-Additive kernels and other contributions in computer experiments

Olivier ROUSTANT

http://www.emse.fr/~roustant

Ecole des Mines de St-Etienne

GDR MASCOT NUM

22th March 2012

Ecole Nationale Supérieure des Mines SAINT-ETIENNE

Outline

1. Background

2. Selected contributions
   Spotlights on metamodeling, SFDs evaluation and software

3. Focus: "Interaction screening"
   Application to the recovery of block-additive structures

Background

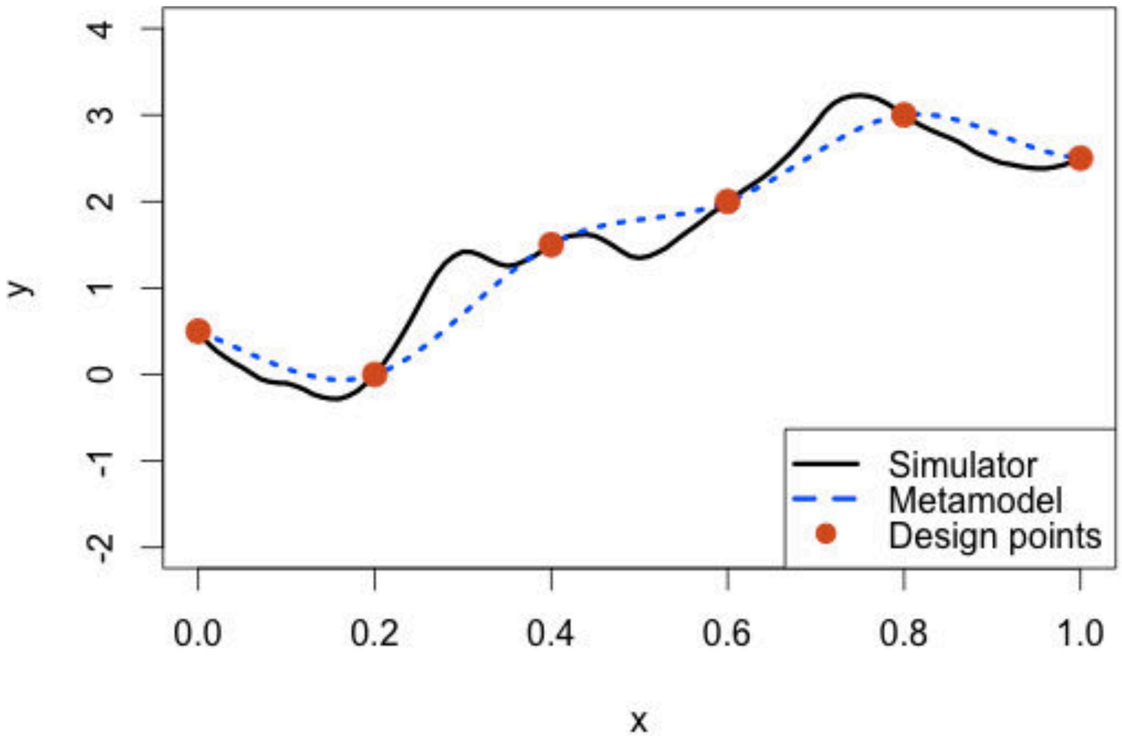
Industrial context

- Time-consuming computer codes
  - car crash-test simulator, thermal hydraulic code in nuclear plants, oil production simulator, etc.

  - xi's : input variables – yj's : the output variables
  - Many possible configurations for the variables: often uncertain, quantitative / qualitative, sometimes spatio-temporal, nested...

Mathematical background

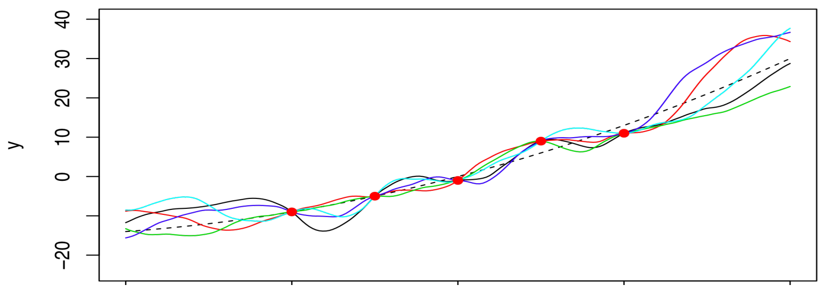
- The idea is to build a metamodel, *computationally efficient*, from a few data obtained with the costly simulator

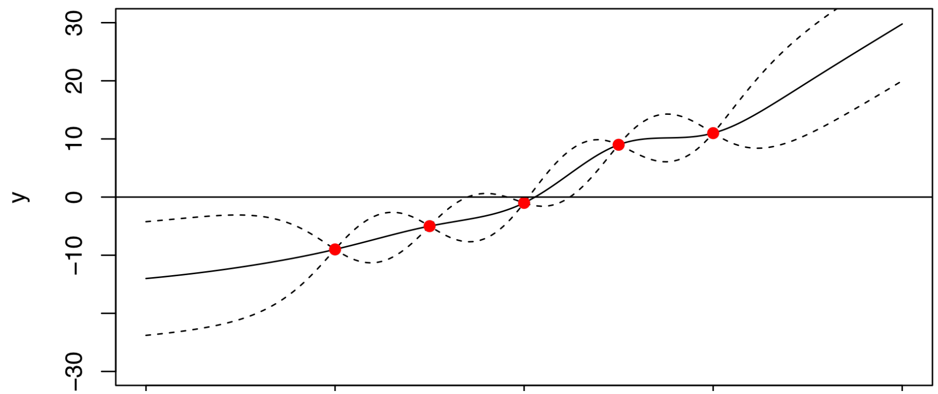


Mathematical background

- **Metamodel building: the probabilistic framework**
  - Interpolation is done by conditioning a Gaussian Process (GP)

    Keywords: GP regression, Kriging model

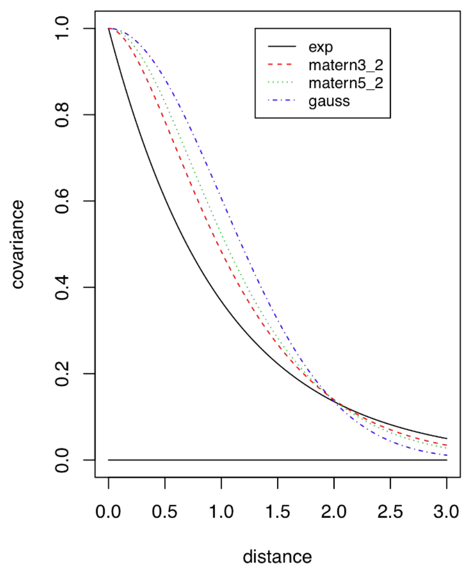


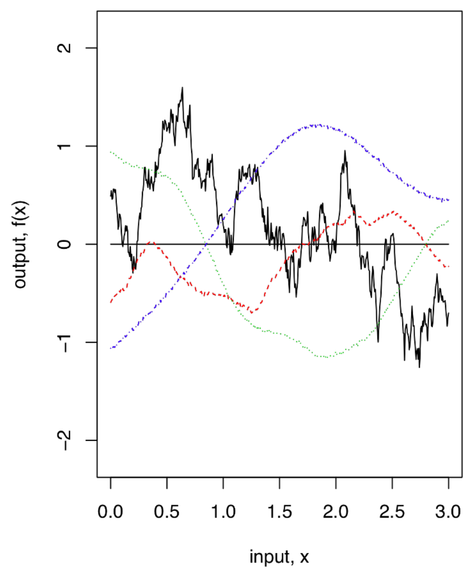


Mathematical background

- Main advantages of probabilistic metamodels:
  - Uncertainty quantification
  - Flexibility w.r.t. the addition of new points
  - Customizable, thanks to the trend and the covariance kernel

$$k(x,x') = cov(Z(x), Z(x'))$$





Smoothness of the sample paths of a stationary process depending on the kernel smoothness at 0

Mathematical background

- Metamodel building: the functional framework
  - Interpolation and approximation problems are solved in the setting of Reproducing Kernel Hilbert Spaces (RKHS), by regularization

- The probabilistic and functional frameworks are not fully equivalent, but translations are possible via the Loève representation theorem

$$\phi : \begin{array}{l} \mathcal{H}_K \to \bar{\mathcal{L}}(Z) \\ K(x,.) \to Z_x \end{array} \qquad \langle K(x,.),\, K(y,.)\rangle = K(x,y) = \langle Z_x,\, Z_y\rangle$$

- In both frameworks, kernels play a key role.

Part 2
Selected contributions

Contributions – Metamodels

Additive kernels

- Additive Kriging [at least: Plate, 1999]
  - Adapt the idea of Additive Models to Kriging

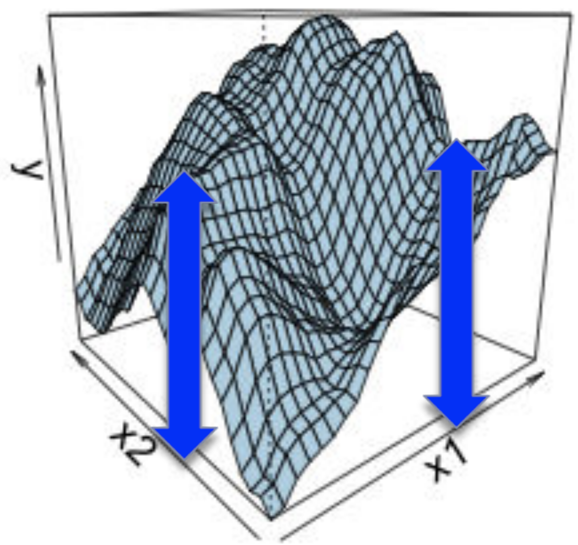
    $Z(x) = Z_1(x_1) + \ldots + Z_d(x_d)$
  - Resulting kernels, for independent processes:

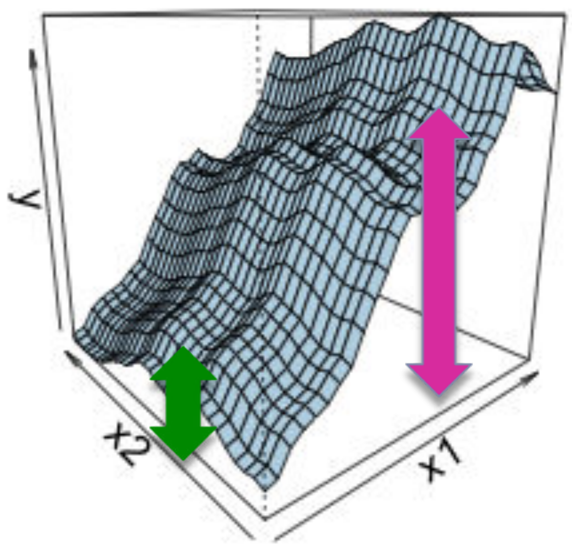
    $$k = k_1 \oplus \ldots \oplus k_d$$
- The aim: To deal with the curse of dimensionality

- Our contribution [Collab with N. Durrande, and D. Ginsbourger]
  - Theory: Equivalence between *kernel* & *sample paths* additivity
  - Empiric: Investigation of a relaxation algorithm for inference

Additive kernels

- Examples of simulations *[package fanovaGraph]*
  - A rigid pattern… with more degrees of freedom





Non-additive kernel
$Z(x) = \sigma Y(x)$

Additive kernel
$Z(x) = \sigma_1 Z_1(x_1) + \sigma_2 Z_2(x_2)$

Block-additive kernels

- The idea [Collab. with T. Muehlenstaedt, J. Fruth, S. Kuhnt and L. Carraro]
  - To identify groups of variables that have no interaction together
  - To use the interactions graph to define block-additive kernels

- New mathematical tools
  - Total interactions
    - Involves the inputs sets containing both x_i and x_j

$$S^{TI}_{\{i,j\}} = \sum_{J \supseteq \{i,j\}} S_J$$

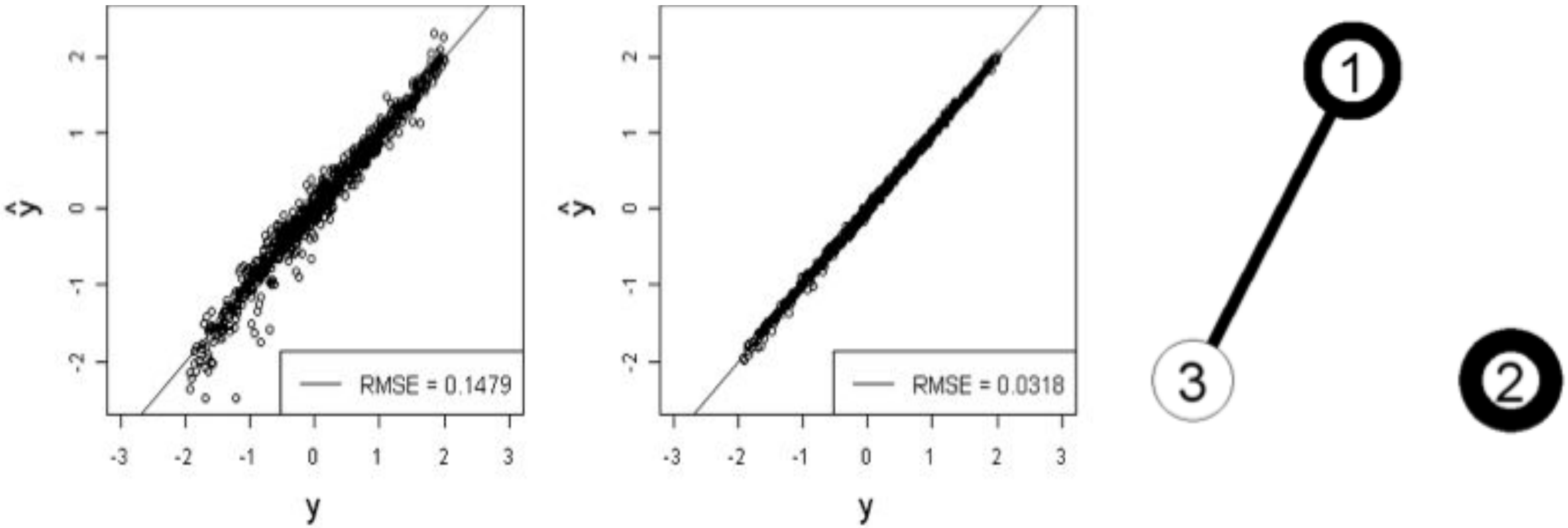
  - FANOVA graph
    - Vertices: input variables – Edges: weighted by the total interactions

Block-additive kernels

- Illustration of the idea relevance on the Ishigami function

  $f(\mathbf{x}) = \sin(x_1) + A\sin^2(x_2) + B(x_3)^4\sin(x_1) = f_2(x_2) + f_{1,3}(x_1,x_3)$

$k = k_1 \otimes k_2 \otimes k_3$ $\qquad$ $k = k_2 \oplus (k_1 \otimes k_3)$

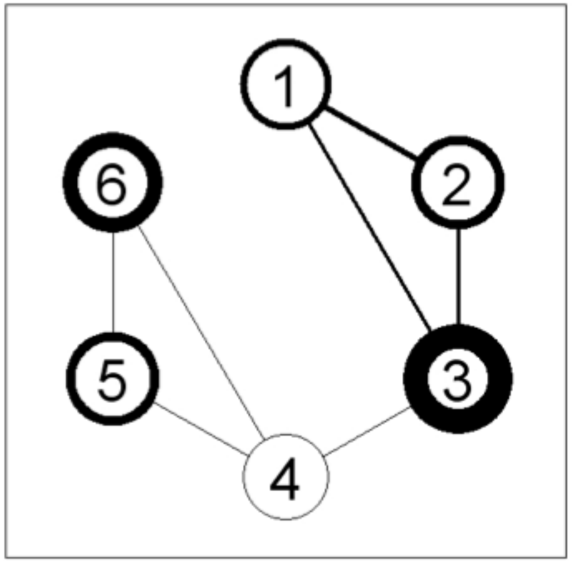


Block-additive kernels

- Illustration of the blocks identification on a 6D function ("b")

$$f(\mathbf{x}) = \cos([1, x_1, x_2, x_3]a') + \sin([1, x_4, x_5, x_6]b') + \tan([1, x_3, x_4]c')$$

ESTIMATION + THRESHOLDING



$$f(\mathbf{x}) = f_{1,2,3}(x_1, x_2, x_3) + f_{4,5,6}(x_4, x_5, x_6) + f_{3,4}(x_3, x_4)$$

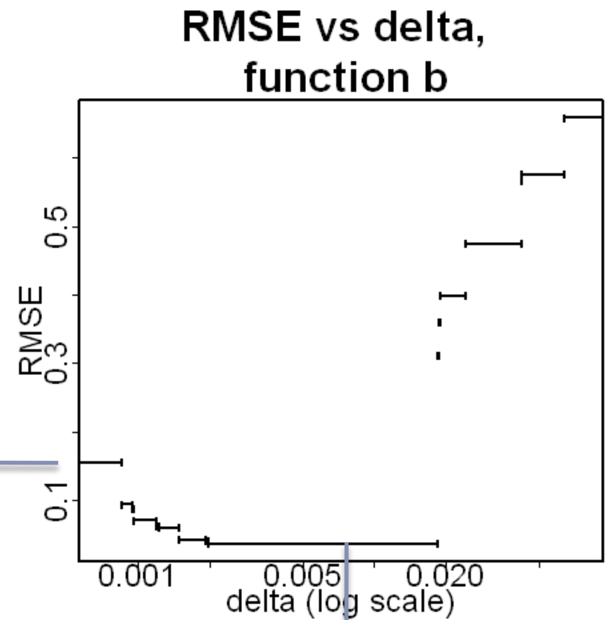
$$Z(\mathbf{x}) = Z_{1,2,3}(x_1, x_2, x_3) + Z_{4,5,6}(x_4, x_5, x_6) + Z_{3,4}(x_3, x_4)$$

Indep. Assump.

$$k(\mathbf{h}) = k_{1,2,3}(h_1, h_2, h_3) + k_{4,5,6}(h_4, h_5, h_6) + k_{3,4}(h_3, h_4)$$

Block-additive kernels

- Graph thresholding issue
  - Sensitivity of the method accuracy to the graph threshold value

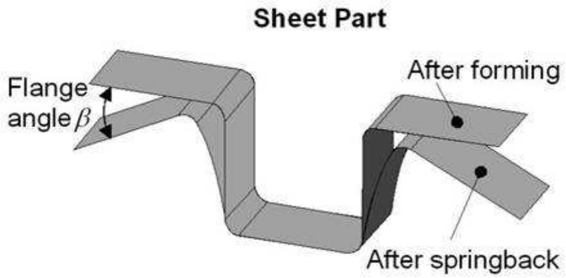


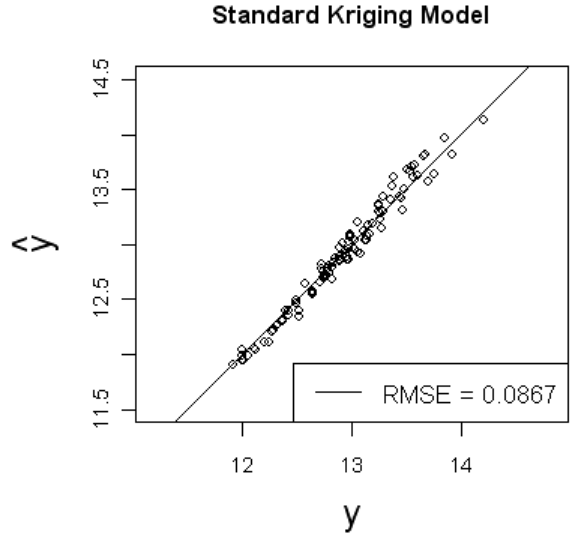
Additive kernel (empty graph)

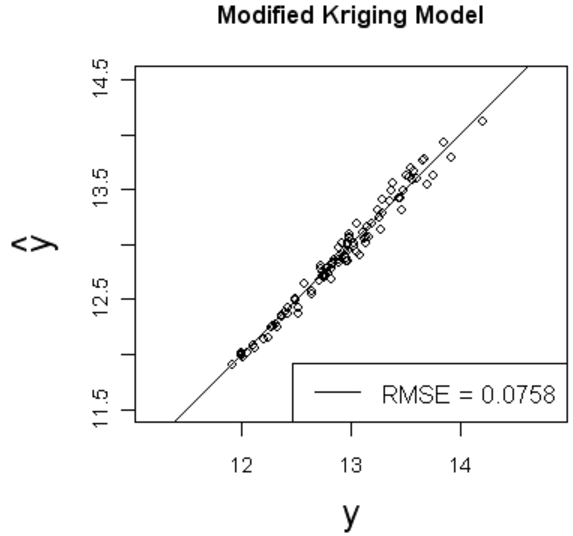
Tensor product kernel (full graph)

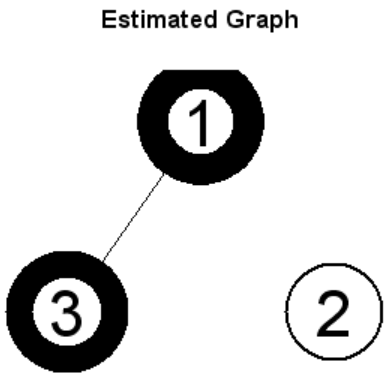
Optimal block-additive kernel

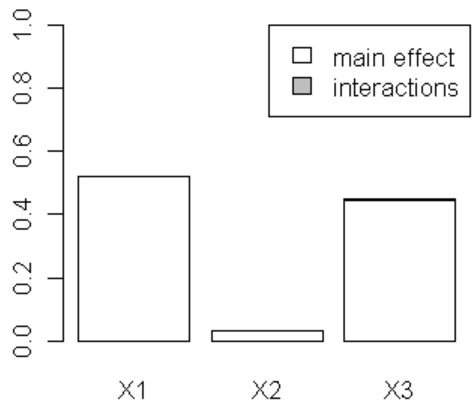
Application to two case studies



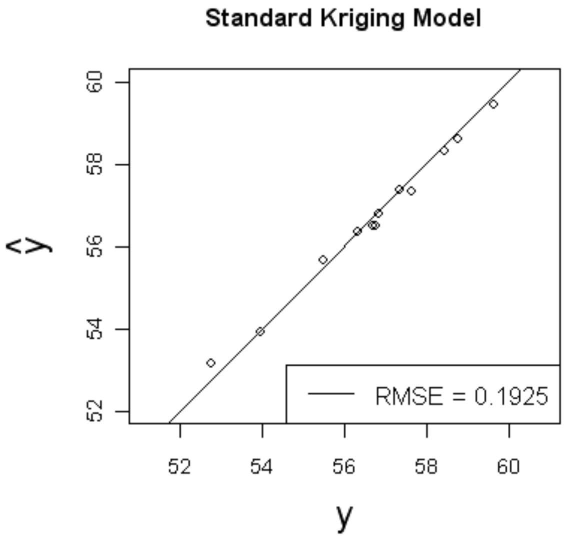


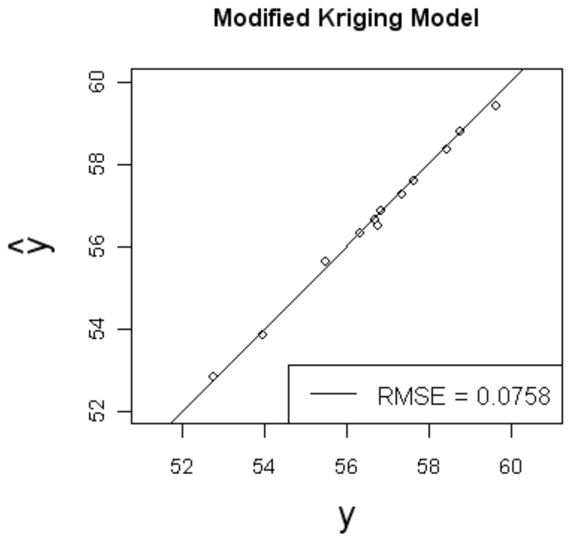






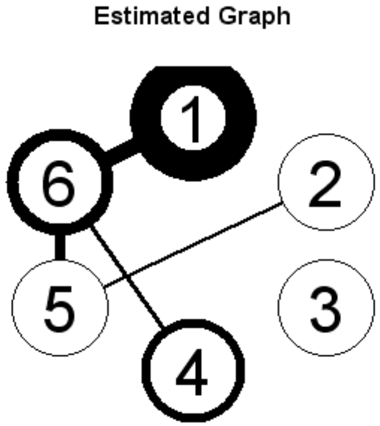
Application to two case studies

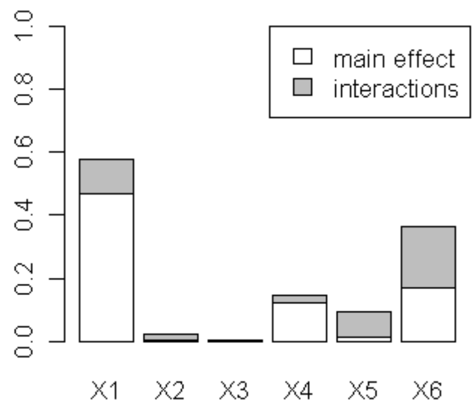




Piston slap problem

Leave-One-Out (RMSE):
0.0864 -> 0.0371





Kernels for Kriging mean SA

- Motivation:
  - To perform a sensitivity analysis (independent inputs) of the proxy
  - To avoid the curse of recursion

- The idea [Collab. with N. Durrande, D. Ginsbourger and L. Carraro]
  - Adapt the ANOVA kernels,

$$k = (1 + k_1) \otimes \ldots \otimes (1 + k_d)$$

    based on the fact that the FANOVA decomposition of

$$f = (1 + f_1) \otimes \ldots \otimes (1 + f_d)$$

    where the f_i's are zero-mean functions, is obtained directly by expanding the product (Sobol, 1993)

Kernels for Kriging mean SA

- Solution with the functional interpretation
  - Start from the 1d- RKHS H_i with kernel k_i
  - Build the RKHS of zero-mean functions in H_i, by considering the linear form L_i: $h \rightarrow \int h(t) d\nu_i(t)$. Its kernel is:

$$k_{i,0}(x,y) = k_i(x,y) - \frac{\int k_i(x,s) d\nu_i(s) \times \int k_i(y,t) d\nu_i(t)}{\iint k_i(s,t) d\nu_i(s) d\nu_i(t)}$$

  - Use the modified FANOVA kernel

$$k = (1 + k_{1,0}) \otimes \ldots \otimes (1 + k_{d,0})$$

Kernels for Kriging mean SA

With this kernel, the Sobol indices **at any order** of the corresp. Kriging mean are computed **analytically without recursion**

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

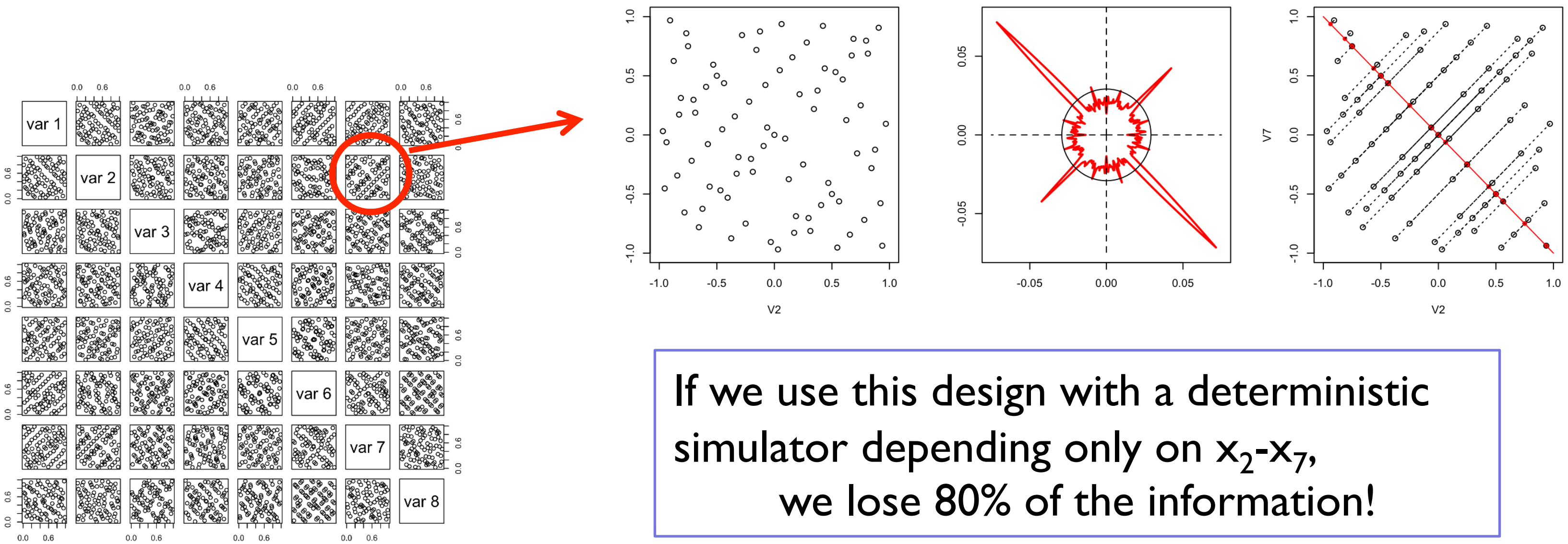
$$S_I = \frac{\mathrm{var}(m_I(\mathbf{X}_I))}{\mathrm{var}(m(\mathbf{X}))} = \frac{\mathbf{F}^t\mathrm{K}^{-1}\left(\bigodot_{i\in I}\Gamma_i\right)\mathrm{K}^{-1}\mathbf{F}}{\mathbf{F}^t\mathrm{K}^{-1}\left(\bigodot_{i=1}^{d}(1_{n\times n}+\Gamma_i)-1_{n\times n}\right)\mathrm{K}^{-1}\mathbf{F}}$$

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.

Contributions – Designs

Selection of an initial design

- The radial scanning statistic (RSS)
  - Automatic defects detection in 2D or 3D subspaces
  - Visualization of defects
  - Underlying mathematics:
    - law of a sum of uniforms, GOF test for uniformity based on spacings



If we use this design with a deterministic simulator depending only on x_2-x_7,
we lose 80% of the information!

Selection of an initial design

- Context: first investigation of a deterministic code
- Two objectives, and the current practice:
  - To catch the code complexity

    space-filling designs (SFDs)
  - To avoid losing information by dimension reduction

    space-fillingness should be stable by projection onto margins

- Our contribution [Collab. with J. Franco, A. Jourdan and L. Carraro]:
  - Dimension reduction techniques involve variables of the form b'x

  space-fillingness should be stable by projection onto oblique straight lines

Selection of an initial design

- Application of the RSS to design selection

Table 1 Worst value of Greenwood statistic for 8-dimensional SFDs of size 80

Design type[a]	Statistic value[b]
Uniform	0.039 (0.003)
Maximin Latin hypercube	0.048
Audze-Eglais Latin hypercube	0.037
Halton sequence	0.244
Faure sequence	0.161
Sobol sequence	0.101
Sobol sequence, with Owen scrambling	0.041 (0.006)
Sobol sequence, with Faure-Tezuka scrambling	0.088 (0.010)
Sobol sequence, with Owen + Faure-Tezuka scrambling	0.041 (0.006)
Strauss	0.040 (0.004)

Contributions – Software

Software for data analysis

- The need
  - To apply the applied mathematics on industrial case studies
  - To investigate the proposed methodologies
  - To re-use our [own!] codes 1 year later (hopefully more)…
- The software form
  - R language:
    - Freeware - Easy to use - Huge choice of updated libraries (packages)
  - User-friendly software prototypes
    - Trade-off between professional quality (unwanted) and un-re-usable codes

Software for data analysis

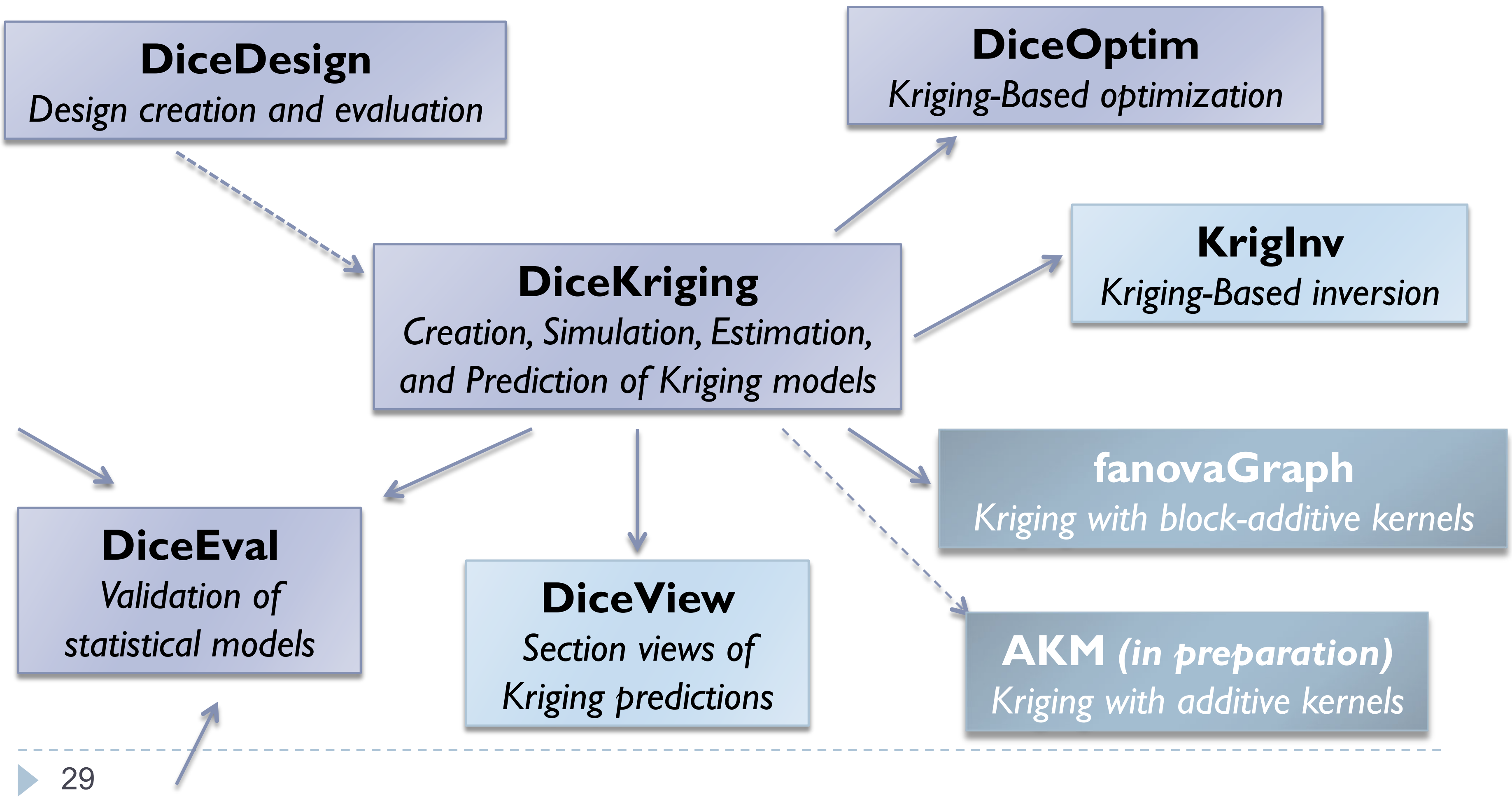
- The packages and their authors
  - A collective work: Supervisors [really], (former) PhD students and… some brave industrial partners!
    - **DiceDesign**: *J. Franco, D. Dupuy, O. Roustant*
    - **DiceKriging**: *O. Roustant, D. Ginsbourger, Y. Deville*
    - **DiceOptim**: *D. Ginsbourger, O. Roustant*
    - **DiceEval**: *D. Dupuy, C. Helbert*

    - **DiceView**: *Y. Richet, Y. Deville, C. Chevalier*
    - **KrigInv**: *V. Picheny, D. Ginsbourger*

    - **fanovaGraph**: *J. Fruth, T. Muehlenstaedt, O. Roustant*
    - ***(in preparation)*** **AKM**: *N. Durrande*

Software for data analysis

- The Dice packages (Feb. and March 2010) and their satellites



Software with data analysis

- Some comments about implementation [ongoing work with D. Ginsbourger, and Y. Deville]
- Leading idea
  - The code should be as close as possible as the underlying maths
    - Example: Operations on kernels.

Illustration with isotropic kernels

$$k_{\text{iso}}(x, y; \theta) = k(x, y; \theta, \ldots, \theta)$$

$$\frac{\partial k_{\text{iso}}}{\partial \theta}(x, y; \theta) = \sum_{i=1}^{p} \frac{\partial k}{\partial \theta_i}(x, y; \theta, \ldots, \theta)$$

Unwanted solution: to create a new program k_{iso} for each new kernel k

Implemented solution: to have the same code for any basis kernel k

Tool: object-oriented programming

Contributions – References

- Additive kernels:
  - PhD thesis of N. Durrande
  - N. Durrande, D. Ginsbourger, O. Roustant (+2012), "*Additive covariance kernels for high-dimensional Gaussian process modeling*", in revision for the *Annales de la Faculté des Sciences de Toulouse*

- Block-Additive kernels:
  - J. Fruth, O. Roustant, S. Kuhnt (+2011), "*Total interaction indices for the decomposition of functions with high complexity*", HAL.
  - T. Muehlenstaedt, O. Roustant, L. Carraro, S. Kuhnt (2011), "*Data-driven Kriging models based on FANOVA-decomposition*", published online in *Statistics & Computing*.

- ANOVA* kernels
  - PhD thesis of N. Durrande (2011)
  - N. Durrande, D. Ginsbourger, O. Roustant, L. Carraro (+2012), "Reproducing kernels for spaces of zero mean functions. Application to sensitivity analysis", in revision for the *Journal of Multivariate Analysis*

- Radial Scanning Statistic:
  - A first version in the PhD thesis of J. Franco (2009)
  - The actual one in: O. Roustant, J. Franco, L. Carraro, A. Jourdan (2010), "*A radial scanning statistic for selecting space-filling designs in computer experiments*", in A. Giovagnoli, A.C. Atkinson, B. Thorsney and C. May, "MODA - 9 - Advances in Model-Oriented Design and Analysis", Springer (Physica-Verlag), p. 189-196

- Software
  - See slide 25 for the packages authors' names
  - O. Roustant, D. Ginsbourger, Y. Deville (+2012), "*DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization*", in revision for the *Journal of Statistical Software.*

- For a synthesis: O. Roustant, mémoire d'HDR, coming soon (on my webpage)

Part 3
Focus: Interaction screening

Ongoing research, in collaboration with J. Fruth and S. Kuhnt

FANOVA-Hoeffding decomposition
(Efron and Stein, 1981, Hoeffding 1948, Sobol later)

- Assume that $X_1, \ldots, X_d$ are independent random variables. Let f be a function defined on D in R^d . Then f is uniquely decomposed as:

$$f(\boldsymbol{X}) = \mu_0 + \sum_{i=1}^{d} \mu_i(X_i) + \sum_{i<j} \mu_{ij}(X_i, X_j) + \cdots + \mu_{1,\ldots,d}(X_1, \ldots, X_d)$$

with the centering conditions:

$$\mathrm{E}(\mu_I(X_I)) = 0, \quad I \subseteq \{1, \ldots, d\}$$

and the non-simplification conditions, implying orthogonality:

$$\mathrm{E}(\mu_{ii'}(X_i X_{i'}) \mid X_i) = \mathrm{E}(\mu_{ii'i''}(X_i X_{i'} X_{i''}) \mid X_i X_{i'}) = \cdots = 0.$$

FANOVA decomposition
(main effects, interactions)

- The terms are obtained recursively:
  - Mean, Main effects

$$\mu_0 = E(f(X)) \qquad \mu_i(X_i) = E(f(X)|X_i) - \mu_0$$

  - 2nd order interactions

$$\mu_{ij}(X_i, X_j) = E(f(X)|X_i, X_j) - \mu_i(X_i) - \mu_j(X_j) - \mu_0$$

  - And more generally:

$$\mu_I(X_I) = E(f(X)|X_I) - \sum_{I' \subsetneq I} \mu_{I'}(X_{I'})$$

FANOVA decomposition
(Sobol indices)

- The name "FANOVA" becomes from the relation on variances implied by orthogonality:

$$D = \mathrm{var}(f(\boldsymbol{X})) = \mathrm{var}(\mu_0) + \sum_{i=1}^{d} \mathrm{var}(\mu_i(X_i)) + \sum_{i<j} \mathrm{var}(\mu_{ij}(X_i, X_j)) + \cdots + \mathrm{var}(\mu_{1,\ldots,d}(X_1, \ldots, X_d))$$

- (unnormalized) Sobol indices:

$$D_I = \mathrm{var}(\mu_I(X_I))$$

FANOVA decomposition
(Total indices)

- The total index of one variable X_i implies all the subsets J containing {i}

$$D_i^T = \sum_{J \supseteq \{i\}} D_J$$

- Extension for a group of variables X_I: implies all the subsets J that contain at least one element in I (or equivalenty, that are not contained in – I)

$$D_I^T = \sum_{\substack{J \\ J \cap I \neq \emptyset}} D_J$$

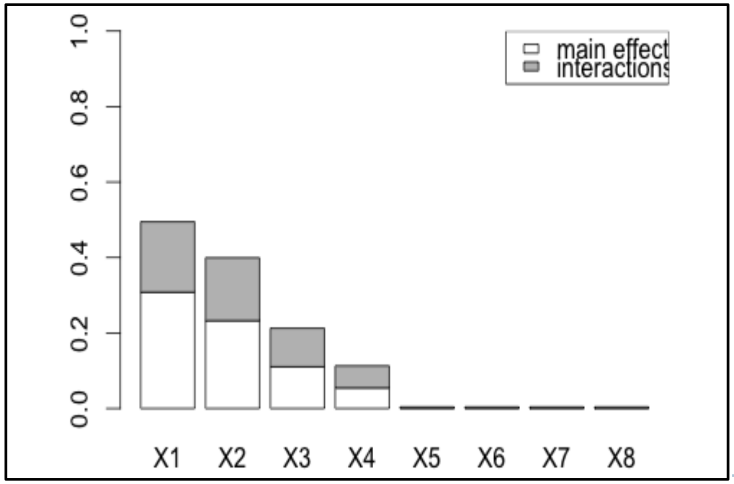
Total indices and screening

- If $D_i^T = 0$, the variable X_i is removed (no terms containing X_i)
  - Remark: A condition is required on the probability measure

$$g(X_1, \ldots, X_d) = \prod_{k=1}^{d} \frac{|4X_k - 2| + a_k}{1 + a_k}$$

a = (0, 1, 4.5, 9, 99, 99, 99, 99)

[package sensitivity]



Total indices of the g-Sobol function:

X_5, X_6, X_7, X_8
can be removed

Total interactions & FANOVA graph

- The total interaction index of a group of variables X_I implies all subsets J containing I. For a pair:

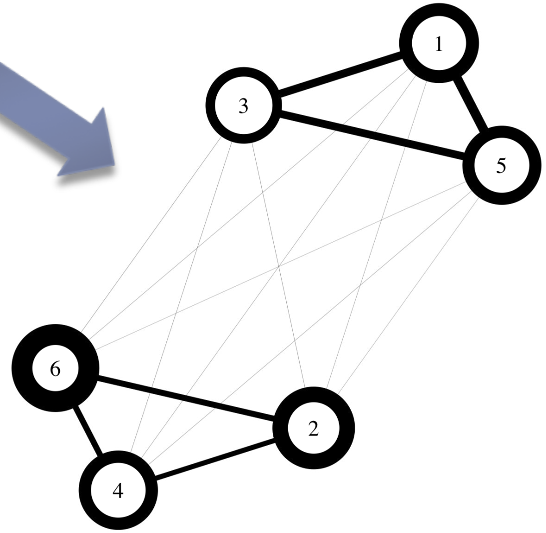
$$D_{i,j}^{TI} = \mathfrak{D}_{ij} := \mathrm{var}\left(\sum_{I \supseteq \{i,j\}} \mu_I(X_I)\right) = \sum_{I \supseteq \{i,j\}} D_I$$

- The FANOVA graph is a valued graph with:
  - Vertices: the input variables (weigth: main effect)
  - Edges: exists if the total interaction index is >0, (weight: its value)

Total interactions & Interaction screening

- If $D_{i,j}^{TI} = 0$, the interaction (X_i, X_j) is removed in the graph (no terms containing both X_i and X_j)
  - Remark: A condition is required on the probability measure

$$f(X_1, \ldots, X_6) = \cos([1, X_1, X_5, X_3]\,\beta) + \sin([1, X_4, X_2, X_6]\,\gamma)$$



Total interaction indices of f:

All the interactions (X_i, X_j) with i in the 1st group {1,3,5} and j in the 2nd one {2,4,6} can be removed

Total interaction indices – Theory

FANOVA decomposition
(Closed indices)

- The closed index of a group of variables X_I implies all subsets J contained in I

$$D_I^C = \mathrm{var}\left(E[f(\boldsymbol{X})|X_I]\right) = \sum_{J \subseteq I} D_J$$

- The link with total indices is the following:

$$D = D^C_{-I} + D_I^T$$

First formula

- There is an obvious link between total interaction indices and total effects of a group of variables

Proposition 1

$$\mathfrak{D}_{ij} = D_i^T + D_j^T - D_{i,j}^T$$

$$\mathfrak{D}_{ij} = D + D^C_{-\{i,j\}} - D^C_{-i} - D^C_{-j}$$

Second formula ("fixing method")

- Fix $x_3, \ldots, x_d$, and consider the 2^{nd} order interaction of the 2-dimensional function:

$$(x_1, x_2) \rightarrow f(\mathbf{x}) = f_1(x_{-2}) + f_2(x_{-1}) + f_{12}(x_1, x_{2;}\ x_{-\{1,2\}})$$

- Denote $D_{12|x3,\ldots,xd}$ the its index. Then, the t. i. index is obtained by integration over $x_3,\ldots,x_d$.

Second formula ("fixing method")

Proposition 2

$$\mathfrak{D}_{ij} = \mathrm{E}\left(D_{i,j|X_{-\{i,j\}}}\right)$$

Proofs (see [Fruth et al., 2011])

1. *With the FANOVA decomposition of the 2-dimensional functions*
2. *Via total indices*

Total interaction indices – Estimation

Estimators of the total interaction indices

- Via closed effects with Monte Carlo (Sobol method)

$$\widehat{\mathfrak{D}}_{ij} = \widehat{D} + \widehat{D}^C_{-\{i,j\}} - \widehat{D}^C_{-i} - \widehat{D}^C_{-j}$$

- Via total effects with RBD-FAST

$$\widehat{\mathfrak{D}}_{ij} = \widehat{D}^T_i + \widehat{D}^T_j - \widehat{D}^T_{\{i,j\}}$$

- FAST + (usual) Monte Carlo, for the fixing method

$$\widehat{\mathfrak{D}}_{ij} = \frac{1}{n_{\mathrm{MC}}} \sum_{k=1}^{n_{\mathrm{MC}}} \widehat{D}^k_{i,j|X_{-\{i,j\}}}$$

Estimators: some properties

Estimator	Positivity	Bias	Variance
Closed effects / Sobol method	NO	0	?
Total effects / RBD-FAST	NO	can be large	?
Fixing method / FAST + MC	YES	small	?

Estimators: numerical cost

- Number of function evaluations to evaluate all the total interaction indices

RBD-FAST	$N = 2(Md + L) \times \left(\binom{d}{2} + d\right)$
Sobol method	$N = n_{Sobol} \times \left(\binom{d}{2} + d + 1\right)$
Fixing method	$N = \binom{d}{2} \times n_{MC} \times n_{FAST}$

- In this table:
  - d=problem dimension, M=6, n_{FAST}=500 (*to satisfy the positivity constraint*), L(>100), n_{Sobol}, n_{MC} are integers.

Empirical tests

- In the following the 3 estimators are compared for a same number of function evaluations

- Example 1: A 6-dimensional complex function
- Example 2: A function with only one 3rd order interaction
- Example 3: A function with 2nd order interactions only

Example 1: A 6-dim. complex function

- Let us consider the 6D g-Sobol function over $[-0,1]^6$

$$g(X_1,\ldots,X_d)=\prod_{k=1}^{d}\frac{|4X_k-2|+a_k}{1+a_k}$$

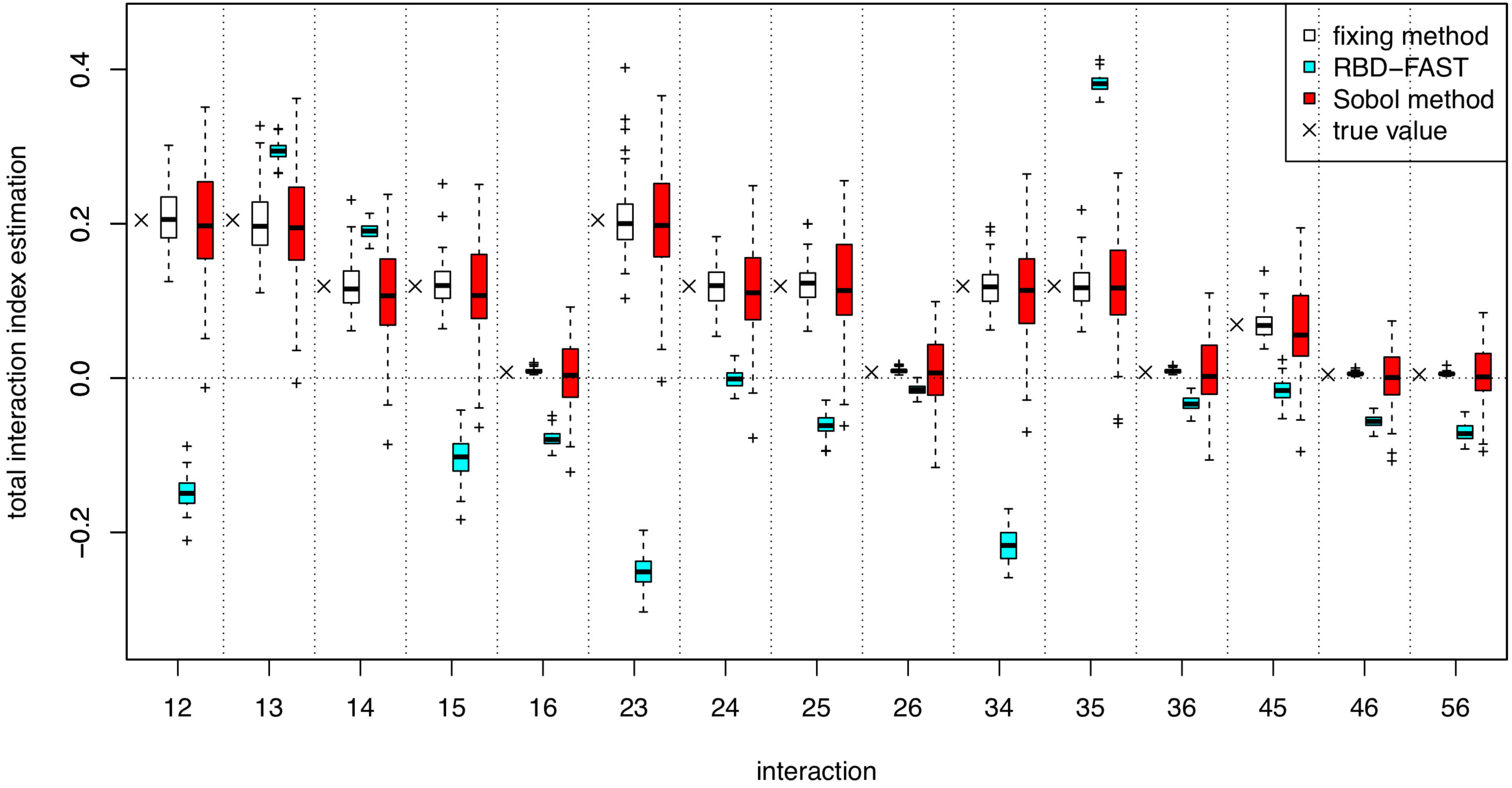
with a = (0, 0, 0, 0.4, 0.4, 5), and uniform distrib.

- This is a complex function:
  - Overall variance: ≈ 3.27
  - Sum of main effects + 2nd order interactions: ≈ 2.06

Example 1: A 6-dim. complex function

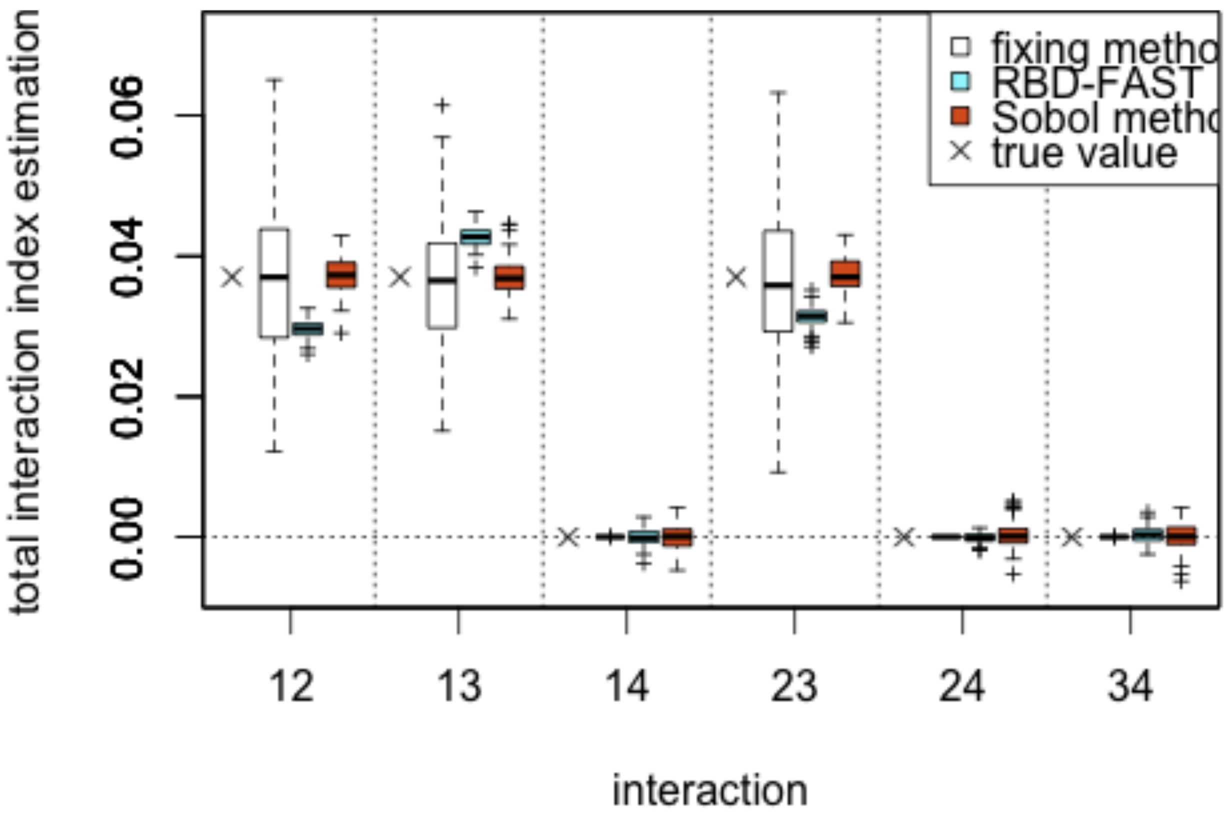
- We compare the three estimators for an equal number of functions evaluations
  - N = 75 000 -> L = 7500, n_{Sobol} = 3 409, n_{MC} = 10
  - N = 600 000 -> 8 times higher values for L, n_{Sobol}, n_{MC}

Example 1: A 6-dim. complex function



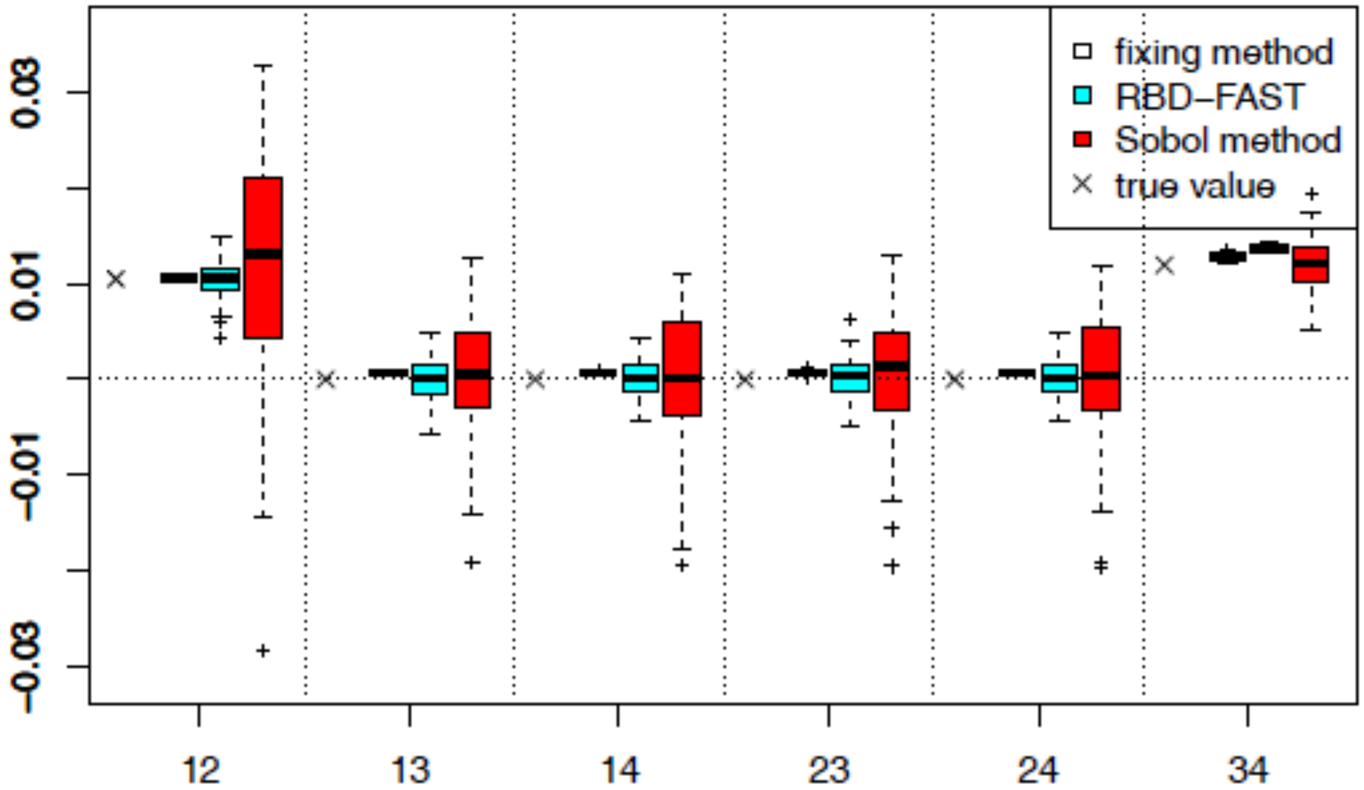
Results for N = 15x40000, obtained with 100 replicates

Example 2: Pure 3rd order interaction



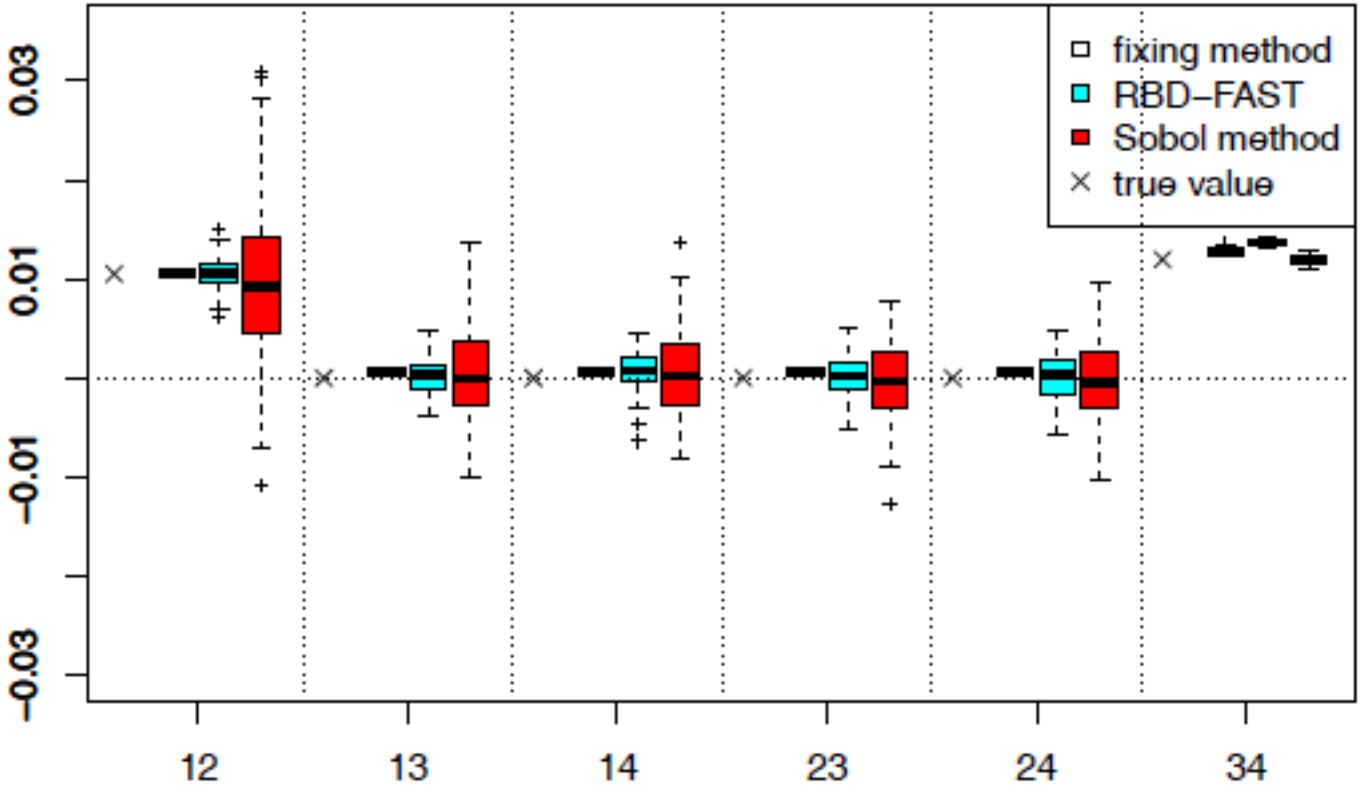
- Results for N = 6x5000, obtained with 100 replicates, for the function: $x_1x_2x_3$, over $[-1,1]^4$ (uniform measure)

Example 3: 2^{nd} order interactions only



- Results for N = 6x5000, obtained with 100 replicates, for the function: $\sin(x_1+x_2) + 0.4*\cos(x_3+x_4)$, over $[-1,1]^4$ (uniform measure)

Example 3: 2^{nd} order interactions only



Remark: With the new estimator for the Sobol method [Janon et al, 2012]

Some important remarks

- The accuracy of the fixing method depends on the variability of the interaction of the fixed function with respect to the fixed variables
  - Very good for second order interaction only
  - Not so good for a (pure) high order interaction
  - Very good when the total interaction is zero

  ➔ Recommended for interaction screening

- RBD-FAST is sometimes highly biased
  - Needs a correction (see [Tissot and Prieur, 2011])

Total interaction indices – Conclusion

Conclusion (1/2)

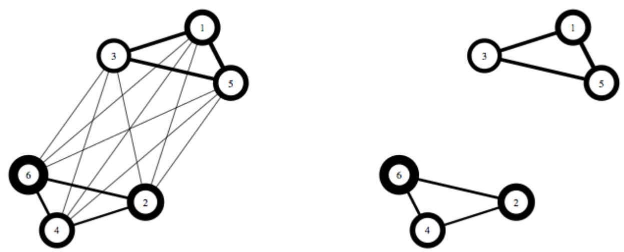
- TII generalizes screening to interactions

- Estimation:
  - The fixing method reduces computations to 2-dim. functions, and is highly accurate to estimate inactive TII.
  - Two other estimators defined over usual estimators for total or closed indices
    - Their accuracy depends on those ones
  - Reasonable global computational cost: $O(d^2)$

Conclusion (2/2)

- Scope:
  - $2 \leq d \leq 20$ (say)
  - Suited to functions with high order interactions
  - Under the assumption "2nd-order interactions only":
    - TII = 2nd order interaction
    - The fixing method is very accurate

- Applications:
  - Data-driven identification of groups of variables
    - Recovery of block-additive structures

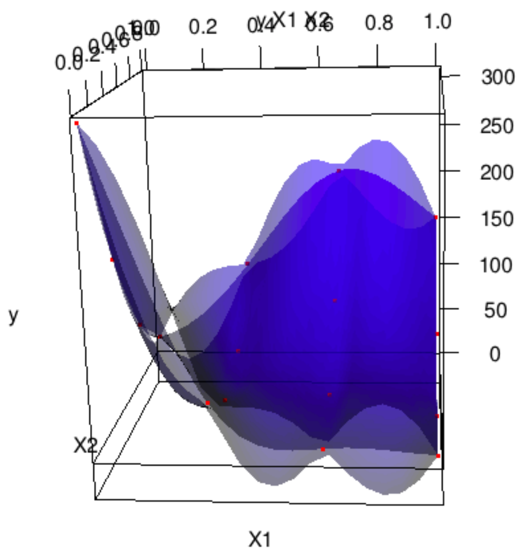


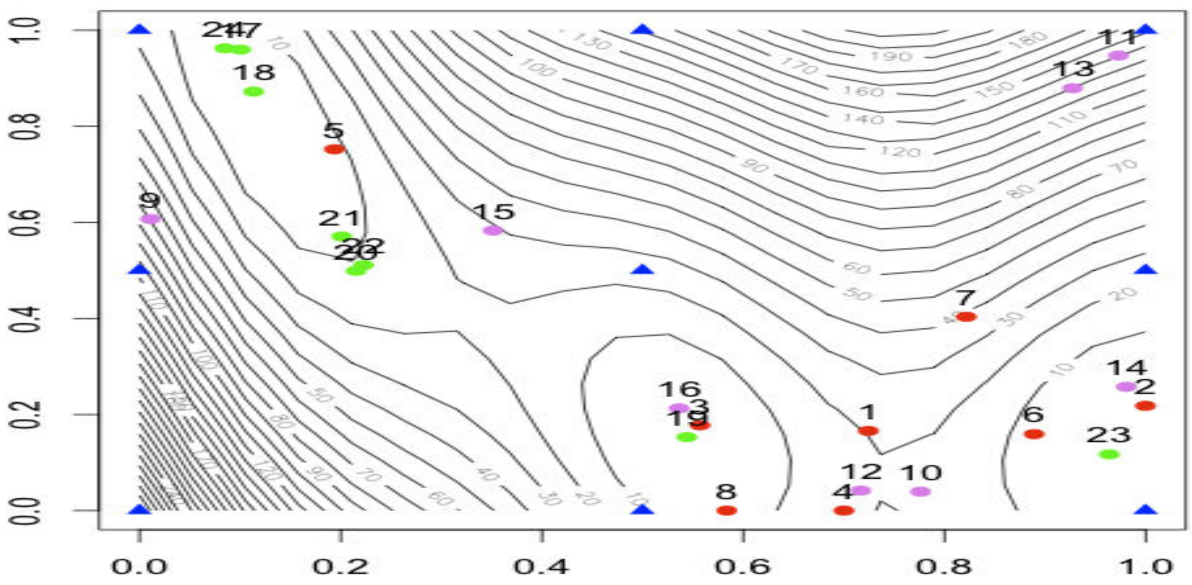
Thank you for your attention!

Supplementary slides

Software for data analysis

- DiceOptim: Kriging-Based optimization
  - Illustration of the adaptive constant liar strategy for 10 processors





Start: 9 points (triangles) – Estimate a Kriging model.
1st stage: 10 points simultaneously (red circles) – Reestimate.
2nd stage: 10 new points simult. (violet circles) – Reestimate.
…

Supplementary slides

- DiceView: 2D (3D) section views of the Kriging curve (surface) and Kriging prediction intervals (surfaces) at a site