MASCOT NUM 2015

8, 9, 10 april

Saint-Étienne

Design and Analysis of Computer Experiments

Invited speakers

G. Allaire, Ecole Polytechnique, France
A. Auger, INRIA - University of Paris-Sud, France
N. Lawrence, University of Sheffield, UK
H. Maruri-Aguilar, Queen Mary, University of London, UK
M. Morris, Iowa State University, USA
J.-S. Park, Chonnam National University, Korea
E. Plischke, Technical University Clausthal, Germany
R. Wilkinson, University of Nottingham, UK

http://mascotnum2015.emse.fr

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINES Saint-Étienne

Groupement de Recherche MASCOT-NUM
Méthodes d'Analyse Stochastique pour les COdes et Traitements NUMériques

Math

ReDICE

cnrs

Day 1: PhD Day – Wednesday April 8th, 2015

9:00		*Registration*	
		Session 1 – Chairman: N. Durrande	
9:45	P. Ray	Opening talk by the Director of Mines St-Étienne	
10:00	Prof. N. Lawrence	Deep Gaussian Processes	
11:00		*Coffee break*	
		Session 2 – Chairman: D. Ginsbourger	
11:20	H. Maatouk	Correspondence between Gaussian processes with inequality constraints and constrained splines	p. 32
11:50	V. Moutoussamy	Excess probability and quantile estimation for monotone codes	p. 40
12:20	C. Vergé	Island particle algorithms and their application to rare event estimation	p. 50
12:50		*Lunch break and poster session*	
		Session 3 – Chairwoman: C. Prieur	
15:00	J. Guerra	Robust construction of a spatio-temporal surrogate model - Application in thermal engineering	p. 22
15:30	C.V. Mai	Polynomial chaos expansions for time-dependent problems	p. 34
16:00	L. Josset	Functional error modeling for Bayesian inference in hydrogeology	p. 26
16:30		*Coffee break*	
		Session 4 – Chairman: V. Picheny	
16:50	G. Damblin	Adaptive numerical designs for the calibration of computer models	p. 12
17:20	M. Binois	High-dimensional Bayesian multi-objective optimization with random embeddings	p. 8

Day 2 – Thursday April 9th, 2015

8:30		*Registration*
		Session 5 – Chairman: O. Roustant
9:00		Conference opening
9:15	Prof. M. Morris	Design and analysis of computer experiments with functional inputs: Sensitivity Analysis
10:15		Mascot-num news
10:40		*Coffee break*
		Session 6 – Chairman: L. Pronzato
11:00	Prof. H. Maruri-Aguilar	Sequential generation of computer experiments
12:00	Prof. G. Allaire	A linearized approach to worst-case design for structural optimization with uncertainties
13:00		*Lunch break and poster session*
		Session 7 – Chairman: ??
14:30	Prof. J-S Park	Algorithms for tuning computer codes, including an EM approach
15:30	Prof. R. Wilkinson	GP-ABC: accelerating inference for intractable stochastic computer models
16:30		*Coffee break*
		Session 8 – Chairman: R. Le Riche
16:50	Prof. A. Auger	Numerical Optimization with CMA-ES: From theory to practice and from practice to theory
19:00		Social event: Guided visit of the Design Biennial
20:00		Gala dinner at the Design Biennial

Day 3 – Friday April 10th, 2015

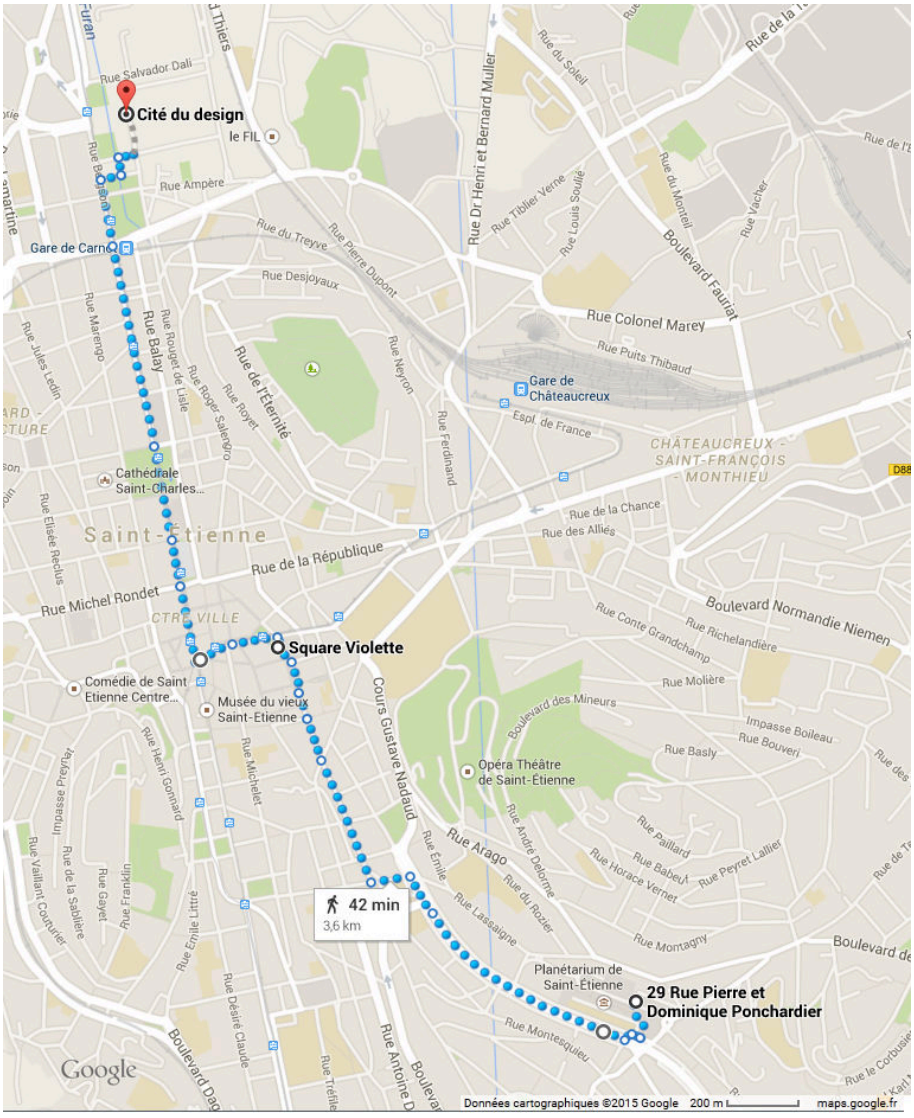
		Session 9
9:30	Prof. M. Morris	Design and analysis of computer experiments with functional inputs and/or outputs: Metamodels
10:30		The ReDICE project: Outcomes and news
11:00		*Coffee break*
		Session 10
11:20	Prof. E. Plischke	Estimating Global Sensitivity Measures: Torturing the Data Until They Confess
12:20		Closure
12:30		*Lunch and departure*

Poster Session

Guided visit of the design Biennal - Gala dinner

The social events will take place at the Cité du Design, on Thursday April 9th, 2015 at respectively 19:00 and 20:00. The Cité du Design is located 3 Rue Javelin Pagnon. It takes around 35 minutes using public transports:

- Take the bus line 6 at *Centre Congrès* located on the main road (Cours Fauriel) until the terminus *Square Violette*.
- Take Tram T2 at station *Peuple Libération* (direction *Terrasse* or *Hôpital Nord*) until stop *Cité du Design*.



Simulating excursion sets under a Gaussian random field prior; how to choose simulation points?

D. Azzimonti
University of Bern

Supervisor(s): PD Dr. David Ginsbourger and Prof. Ilya Molchanov (University of Bern)

Ph.D. expected duration: 2013-2016

Address: University of Bern
Institute of Mathematical Statistics and Actuarial Science
Alpeneggstrasse 22
CH-3012 Bern

Email: dario.azzimonti@stat.unibe.ch

Abstract:

The main contribution of this work is a method to generate conditional realizations of the random excursion set of a Gaussian random field (GRF) based on simulations of the field at few locations. In particular we focus on the problem of estimating the excursion set of a function under a limited evaluation budget. We consider a function $f : D \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ where D is a compact of $\mathbb{R}^d$, $d \geq 1$. f is regarded as a realization of a Gaussian random field with continuous paths, $Z = (Z_{\mathbf{x}})_{\mathbf{x}\in D}$, whose mean function and covariance kernel are denoted by m and k. The object of interest is the set $\Gamma^\star = f^{-1}(T) = \{\mathbf{x} \in D : f(\mathbf{x}) \in T\}$, with $T \in \mathcal{B}(\mathbb{R})$, an element of the Borel σ-algebra of $\mathbb{R}$. Here we assume that T is a closed set in $\mathbb{R}$ of the form $[t, \infty)$ for some $t \in \mathbb{R}$. Figure 1 shows a two dimensional example: the excursion set of the negative Branin-Hoo function f above $t = -10$. Under such assumptions, the excursion set $\Gamma^\star$ is closed in D and $\Gamma = \{\mathbf{x} \in D : Z(\mathbf{x}) \in T\}$ defines a random closed set.

Taking a standard matrix decomposition approach for GRF simulations, a straightforward way to obtain realizations of Γ is to simulate Z at a fine design $G = \{\mathbf{u}_1, \ldots, \mathbf{u}_r\} \subset D$ with large $r \in \mathbf{N}$, and then to represent Γ as $\{\mathbf{u} \in G : Z_{\mathbf{u}} \in T\}$, for example the realizations shown in Figure 1 were obtained with simulations on a 50×50 grid ($r = 2500$). A drawback of this procedure, though, is that it may become impractical for a high resolution r, as the covariance matrix involved may rapidly become close to singular and cumbersome, if not impossible, to store. Alternative GRF simulation techniques could solve the issue under limited assumptions or by relying on approximate solutions, however they were not considered here to preserve the generality of the method.

The proposed approach consists in replacing the conditional GRF simulations at the design G with approximate simulations relying on a smaller simulation design $E_m = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$, with $m \ll r$. In particular we replace Z with the simpler random field $\widetilde{Z}$, which is constructed in such a way that the associated excursion set $\widetilde{\Gamma}$ is as close as possible to Γ. We consider two sets "close" if they have a small expected distance in measure, defined as $d(\Gamma, \widetilde{\Gamma}) = E[|\Gamma \Delta \widetilde{\Gamma}|]$, where $|\cdot|$ indicates the volume of a set and $\Gamma \Delta \widetilde{\Gamma}$ is the symmetric difference between the two sets, see e.g. [3]. Figure 2 shows realizations of $\widetilde{\Gamma}$ reconstructed from simulations at $m = 100$ optimally selected points on the same two dimensional example used previously. Denote by $Z(E_m) = (Z_{\mathbf{e}_1}, \ldots, Z_{\mathbf{e}_m})^T$ the random vector of values of Z at E_m, the essence of the proposed approach is to appeal to affine predictors of Z, i.e. to consider $\widetilde{Z}$ of the form

$$\widetilde{Z}(\mathbf{x}) = a(\mathbf{x}) + \mathbf{b}^T(\mathbf{x}) Z(E_m) \qquad (\mathbf{x} \in D), \tag{1}$$

where $a : D \longrightarrow \mathbb{R}$ is a trend function and $\mathbf{b} : D \longrightarrow \mathbb{R}^m$ is a vector-valued function of deterministic weights.

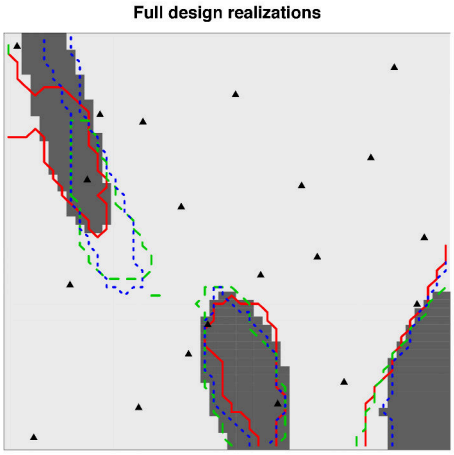


Figure 1: GRF model for the negative Branin-Hoo function with 20 evaluation points (black triangles), the true excursion set (gray shaded area), 3 realizations of the excursion set (colored contours). The realizations are obtained with simulations at a grid 50×50.

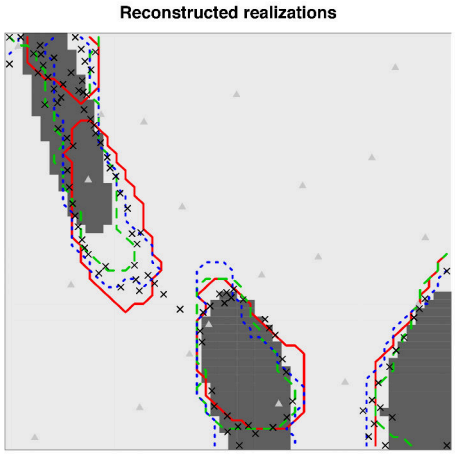


Figure 2: GRF model for the negative Branin-Hoo function with 20 evaluation points (gray triangles). 3 realizations (colored contours) of the random set are generated by simulating at 100 optimally-chosen points (black crosses) and predicting the field at the 50×50 grid.

By using the structure of $\widetilde{Z}$ outlined in Equation 1 we show that it is possible to express the expected distance in measure as a function of the probability distribution of the underlying field only. This property of the expected distance in measure allows us to minimize it as a function of E_m, thus finding the optimal simulation points.

Finally we present two applications that benefit from the optimized selection of simulation points. In the first application we show that the distance average [1] of a set over a two dimensional grid can be computed accurately by simulating the field at few simulation points and by predicting the field over an arbitrary design. In the second application we apply the proposed method to the problem of estimating the distribution of the excursion volume in a six dimensional example.

Joint work with: Julien Bect, Clément Chevalier, David Ginsbourger.

References

[1] A. J. Baddeley and I. S. Molchanov. Averaging of random sets based on their distance functions. *J. Math. Imaging Vision*, 8:79–92, 1998.

[2] C. Chevalier, D. Ginsbourger, J. Bect, and I. Molchanov. Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviation with Gaussian process models. In *mODa 10 Advances in Model-Oriented Design and Analysis*. Physica-Verlag HD, 2013.

[3] I. Molchanov. *Theory of Random Sets*. Springer, 2005.

Short biography – Dario Azzimonti completed his Master degree in Mathematics at the Università degli Studi in Milan in 2012, spending one year at École des Mines de Paris. After a short period in the industry, he started his PhD at the University of Bern in September 2013. His research topics include Gaussian random fields, set estimation, kriging-based inversion and random closed set theory.

High-dimensional Bayesian multi-objective optimization with random embeddings

M. Binois
Mines de Saint-Étienne - Renault

Supervisor(s): D. Ginsbourger (University of Bern) and F. Mercier (Renault) and O. Roustant (Mines de Saint-Étienne)

Ph.D. expected duration: 01/2013 - 12/2015

Adress: 158 Cours Fauriel, CS 62362 42023 Saint-Étienne Cedex 2, France

Email: mickael.binois@mines-stetienne.fr

Abstract:

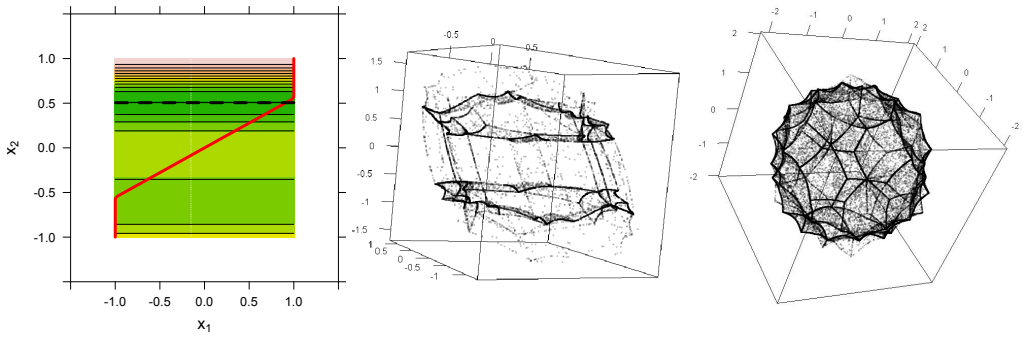
Specifications in the automotive industry are increasingly stringent and compromises must be made between conflicting objectives, leading to multi-objective optimization (MOO). The conception of the various parts of a vehicle relies on accurate but demanding black-box computer codes, resulting in limited budgets of evaluation. This makes the use of surrogate-based methods popular, and particularly those relying on Gaussian Process (GP) modelling [3], a.k.a. kriging. Those metamodels are used to define and optimize acquisition functions such the Expected Improvement (EI) criterion [2] and its multi-objective counterparts, see e.g. [4], to sequentially add new points. We present results towards making those MOO algorithms cope with high-dimensional search spaces under the hypothesis that only a few number d_e of variables are actually influential. To this end, we resort to variations of the recent REMBO (Random EMbedding Bayesian Optimization) method [5] and improve both its robustness and its range of applicability.

Denote $\mathcal{X} \subset \mathbb{R}^D$ the input domain and $f : \mathcal{X} \to \mathbb{R}$ the function to optimize. The principle in a nutshell of the standard REMBO algorithm [5] is to map a smaller-dimensional domain $\mathcal{Y} \subset \mathbb{R}^d$ onto $\mathcal{X}$ using a random matrix $\mathbf{A} \in \mathbb{R}^{D\times d}$ with independent $\mathcal{N}(0,1)$ entries, where $d_e \leq d \ll D$. Given that $\mathcal{Y}$ is large enough, it contains a point reaching the optimum value of f. Box constraints, i.e. $\mathcal{X} = [-1,1]^D$ (eventually obtained via rescaling) are enforced with the convex projection onto $\mathcal{X}$: $p_{\mathcal{X}}$. Bayesian Optimization is applied, e.g. with EI, to optimize the low-dimensional function $g : \mathcal{Y} \to \mathbb{R}$, $g(\mathbf{y}) = f(\mathbf{u}(\mathbf{y}))$ with $\mathbf{u} : \mathcal{Y} \to \mathcal{X}$, $\mathbf{u}(\mathbf{y}) = p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$. The idea is illustrated on Figure 1)a) with $d = 1$ and $D = 2$: the search of the optimum is restricted to the red line instead of the whole domain. Two Gaussian kernels with length scales l were proposed to build GP models: the first one is $k(\mathbf{y},\mathbf{y}') = \exp\left(-\|\mathbf{y}-\mathbf{y}'\|_d^2/2l^2\right)$ and the second one, using the non-linear mapping (or *warping* [3]) $\mathbf{u}$, is $k(\mathbf{y},\mathbf{y}') = \exp(-\|\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{y}')\|_D^2/2l^2)$. They suffer respectively from the non-injectivity of the mapping and from high-dimensional distances. In both cases, it causes difficulties to select the size of $\mathcal{Y}$, implying to resort to strategies such as arbitrarily fixing $\mathcal{Y}$ and splitting the evaluation budget between several random embeddings [5], thus slowing down convergence.

Our contributions are threefold. First we extend REMBO to multi-objective (or constrained) optimization relatively easily since the associated acquisition functions are still mono-objective. Secondly we introduce a new warping: $\Psi_{\mathbf{A}} : \mathcal{Y} \to \mathcal{X}$, $\Psi_{\mathbf{A}}(\mathbf{y}) = \mathbf{z}' + \|p_{\mathcal{X}}(\mathbf{A}\mathbf{y}) - \mathbf{z}'\|_D . \frac{\mathbf{z}'}{\|\mathbf{z}'\|_D}$ with $\mathbf{z}' = (\max(1, \max_{i=1,\dots,D}|z_i|))^{-1}\mathbf{z}$, $\mathbf{z} = p_{\mathbf{A}}(p_{\mathcal{X}}(\mathbf{A}\mathbf{y}))$ where $p_{\mathbf{A}}$ denotes the orthogonal projection onto $\mathbf{A}$ [1]. The aim is that warped points are contained in the low-dimensional subspace spanned by $\mathbf{A}$, Ran($\mathbf{A}$), while preventing points distant in $\mathcal{X}$ to be too close on Ran($\mathbf{A}$) using a distortion with respect to the origin. This improves significantly the robustness of the optimization with only one embedding. Thirdly, we propose further to modify the matrix $\mathbf{A}$ itself. We start by normalizing its rows $\mathbf{A}_i, 1 \leq i \leq D$, to get points on the d-dimensional sphere $\mathbf{S}$. The modified matrix $\widetilde{\mathbf{A}}$ is obtained as a solution of: $\max_{\mathbf{x}_1,\dots\mathbf{x}_D\in\mathbf{S}} \min_{1\leq j<k\leq D} \|\mathbf{x}_j - \mathbf{x}_k\|_d$, with a local optimization

of the normalized $\mathbf{A}$. Figure 1)b-c) are obtained by successively sampling uniformly points from $\mathcal{Y}$ with large bounds, applying $\Psi_{\mathbf{A}}$ and then $\mathbf{A}^{-1}$, the pseudo-inverse of $\mathbf{A}$, to get and display their coordinates on Ran($\mathbf{A}$). This illustrates the effect of the proposed modification: pre-images of vertices of the D-hypercube belonging to the embedding, appearing as peaks, are spread more regularly. It alleviates possibly wrong side effects of distant D-vertices having close pre-images altogether with enhancing optimization results, as indicated by preliminary results. As a by-product, this makes selecting the size of $\mathcal{Y}$ easier as well as optimizing the acquisition function. Finally, we propose to identify when the embedding is probably not suited to the problem at hand.

Applications on classical mono/multi objective benchmark functions and on a test-case from the automotive industry are proposed. The latter corresponds to the rear shock absorber of a car, with 47 parameters, impacted by a block in four load cases. The objectives to minimize are the penetration of the block in the different scenarii along with the mass of the device.



(a) Principle of REMBO (b) $\mathbf{A}^{-1}\Psi_{\mathbf{A}}(\mathcal{Y})$ (c) $\widetilde{\mathbf{A}}^{-1}\Psi_{\widetilde{\mathbf{A}}}(\mathcal{Y})$

Figure 1: Left: contour plot of a bi-variable function depending on x_2 only, the optimum value is reached on the black dotted line. The red line is the embedding of a 1-dimensional subspace in this 2-dimensional domain. Center and right: boundaries of the image of $\mathcal{Y}$ by $\mathbf{A}^{-1}\Psi_{\mathbf{A}}$, informally denoted $\mathbf{A}^{-1}\Psi_{\mathbf{A}}(\mathcal{Y})$, with $d = 3$, $D = 9$, for a random $\mathbf{A}$ and its modification, $\widetilde{\mathbf{A}}$, respectively.

References

[1] M. Binois, D. Ginsbourger, and O. Roustant. A warped kernel improving robustness in Bayesian optimization via random embeddings. *Proceedings of the International Conference on Learning and Intelligent Optimization (LION 9), to appear*, 2015.

[2] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

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

[4] T. Wagner, M. Emmerich, A. Deutz, and W. Ponweiser. On expected-improvement criteria for model-based multi-objective optimization. *Parallel Problem Solving from Nature–PPSN XI*, pages 718–727, 2010.

[5] Z. Wang, M. Zoghi, F. Hutter, D. Matheson, and N. de Freitas. Bayesian optimization in a billion dimensions via random embeddings. *IJCAI*, 2013.

Short biography – After a dual master's degree in engineering and applied mathematics from Mines Saint-Étienne, Mickaël Binois is continuing with a PhD thesis since January 2013. It is funded by Renault under a CIFRE contract, within the ReDICE consortium.

Estimation of probability of detection curves through surrogate models - Application to flaws detection by non destructive test process

BROWNE Thomas
Université Paris Descartes and EDF R&D

Supervisor(s): Pr JC. FORT (Université Paris 5), Dr B. IOOSS (EDF R&D Chatou) and Dr L. LE GRATIET(EDF R&D Chatou)

Ph.D. expected duration: 2014-2017

Adress: EDF R&D Chatou, 6 quai Watier, 78401 Chatou

Email: thomas.browne@edf.fr

Abstract:

EDF carries out Eddy Current Non Destructive Examination in order to ensure integrity of steam generators tubes [2] . Probability of Detection (POD) curves is a standard tool to evaluate the performance of Non Destructive Testing (NDT) procedures [1]. The goal is to assess the quantification of inspection capability for the detection of harmful flaws for the inspected structure. Here is our framework :

- $Y \in \mathbb{R}$: measure of the signal (or its projection) obtained from a NDT procedure, after it has been through the structure to check.
- $a > 0$: length of the flaw (mm) - parameter of interest.
- $X \in \mathbb{R}^d$: all the other influent parameters of the output Y, dependent on either the structure or the NDT process itself, such as its conductivity, permeability. It is considered to be a random vector whose marginal distributions are independent, following the probability density fonction : $(X_1, ..., X_d) \sim (f_1, ..., f_d)$.
- $\delta \in \mathbb{R}$: the noise observation on Y due to uncotroled properties of the inspected structure (its microstructure for instance...) whose pdf (resp. its cdf) is f_δ (resp. F_δ). In a regression model, δ also includes the model error.
- t_s : the signal threshold from which we consider the flaw to be detected, which is to say when $S(a, X, \delta) > t_s$.

Since the whole environment is random, it is clear that given two identical flaws (ie same length a) on two different structures, it is likely to detect one and not the other. One introduces the idea of a probability of detection. Theoretically speaking, the POD is a one dimensional curve whose expression is given by :

$$\begin{aligned} \forall a > 0 \quad \mathrm{POD}(a) &= \mathbb{P}\left(Y(a, X, \delta) > t_s\right) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} \mathbf{1}_{Y(a,x,v)>t_s} \left(f_1 \bigotimes ... \bigotimes f_d\right)(x_1, ..., x_d)\, f_\delta(v) dv dx \end{aligned}$$

However, high costs of the implementation of experimental POD campaigns combined with continuous increase in the complexity of configuration make them sometimes unaffordable. To overcome this problem, it is possible to resort to numerical simulation of NDT process. For this paper we used Code_Carmel3D developed by EDF R&D (high-time consuming : several hours per computation). Then regression models on the signal output Y makes it possible to build an estimator of the curve POD. To this effect we will introduce three different regression models : linear regression, Gaussian process regression and Gaussian process regression with noisy observations.

Let us assume that Y is an increasing function of both a and δ. Therefore if we call $Y_{a,X}(.)$ the function that evaluates the signal for any noise δ knowing a and X, we can consider its reciprocal

function $Y_{a,X}^{-1}(.)$. Then :

$$\forall a > 0 \quad \text{POD}(a) \quad = \mathbb{P}\left(\delta > Y_{a,X}^{-1}(t_s)\right)$$

which means that the POD can be considered as a cdf itself. From this result it is obvious that the distribution of δ is going to be an important point in our work. A major expectation from this Phd work is to build a tool that evaluates the influence of the parameters X on the POD. So far the randomness of X and δ are put on the same level and any estimation of the POD will be based on assumptions made on F_δ. We modify our framework : let's now consider the random cdf π_X defined as follows

$$\begin{aligned} \forall a > 0 \quad \pi_X(a) \quad &= \mathbb{P}\left(Y(a, X, \delta) > t_s \mid X\right) \\ &= 1 - F_\delta\left(Y_{a,X}^{-1}(t_s)\right) \quad \text{knowing } X \end{aligned}$$

which is to say that for each realization $x \in \mathbb{R}^d$ of X we have a matching cdf which happens to be the POD under the known parameters x. The goal is now to estimate the "distribution" of the random cdf π_X from a collection of its realizations $(\pi_x)_{x \text{ realization of } X}$. Let us set a new framework to compare the cdf by using the 2-Wasserstein distance which is for two cdf F and G :

$$W_2^2(F, G) = \int_0^1 \left(F^{-1}(u) - G^{-1}(u)\right)^2 du$$

where F^{-1} and G^{-1} are actually the quantile functions. Some tools can be defined from this distance to quantify π_X such as a mean (Frechet mean), a variance or quantiles. Their definitions are in fact a natural extension of the real random variables field. Those new definitions for the mean and quantiles (which are cdf functions too in our case) are illustrated in (1).

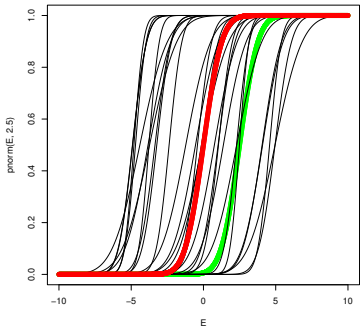


FIGURE 1 – In black, 30 realizations of π_X, in red its "mean" and in green its "75%-quantile". One can notice that in our case the POD-mean and the POD-quantile are cdf as well.

Références

[1] L. Gandosi and C. Annis. Probability of detection curves : Statistical best-practice. *ENIQ TGR Technial Document*, 41, 2010.

[2] L. Maurice, V. Costan, E. Guillot, and P. Thomas. Eddy current NDE performance demonstrations using simulation tools. *Review of Progress in Quantitative Non Destructive Evaluation, Denver, Colorado, USA*, 32 :464–471, 2012.

Short biography – I graduated from Université Paris Descartes in 2014 with a master's degree in probability and statistics . After a 6 month long internship at EDF where I developped a surrogate model for stochastic codes, I moved on a PhD which is mainly about the construction of an estimator for the POD curve.

MascotNum2015 conference - Adaptive numerical designs for the calibration of computer models

Guillaume Damblin
AgroParisTech UMR Inra 518/EDF R&D

Supervisor(s): Prof. Eric Parent (AgroParisTech), Dr. Pierre Barbillon (AgroParisTech) and Dr. Merlin Keller (EDF R&D)

Ph.D. expected duration: 2012-2015

Adress: EDF R&D, 6 quai Watier 78401 Chatou

Email: guillaume.damblin@edf.fr

Abstract:

Making good predictions of a physical system $r(\mathbf{x})$ using a computer model $y_{\boldsymbol{\theta}}(\mathbf{x})$ requires its inputs to be carefully specified. The vector $\mathbf{x}$ of control variables have to mimic physical conditions whereas other inputs $\boldsymbol{\theta}$, called parameters, are specific to the computer model and most often uncertain. Let us suppose that the model error is negligible. Hence,

$$\exists \boldsymbol{\theta} \in \mathcal{T} \ ; \ \forall \mathbf{x} \in \mathcal{X}, \ \ r(\mathbf{x}) = y_{\boldsymbol{\theta}}(\mathbf{x}). \tag{1}$$

The goal of calibration (Kennedy and O'Hagan, 2001; Campbell, 2006) consists in tuning $\boldsymbol{\theta}$ to make the outputs of the computer model as close as possible to the physical measurements $\mathbf{z} = (z_1 = z(\mathbf{x_1}), \cdots, z_n = z(\mathbf{x_n}))$. Assuming the unbiaised framework (1), we have

$$z_i = y_{\boldsymbol{\theta}}(\mathbf{x_i}) + \epsilon_i \tag{2}$$

where $\epsilon_i \sim \mathcal{N}(0, \lambda^2)$ is the measurement error model. Where *prior* informations are available, the calibration consists in conducting a Bayesian inference (Bernardo and Smith, 1994) in order to update the *prior* probability distribution $\Pi(\boldsymbol{\theta})$ yielding the *posterior* distribution $\Pi(\boldsymbol{\theta}|\mathbf{z})$. Under the statistical model (2),

$$\Pi(\boldsymbol{\theta}|\mathbf{z}) \propto \frac{1}{(\sqrt{2\pi}\lambda)^n} \exp\Big[-\frac{1}{2\lambda^2} SS(\boldsymbol{\theta})\Big] \Pi(\boldsymbol{\theta}), \tag{3}$$

where

$$SS(\boldsymbol{\theta}) = ||\mathbf{z} - y_{\boldsymbol{\theta}}(\mathbf{x})||^2. \tag{4}$$

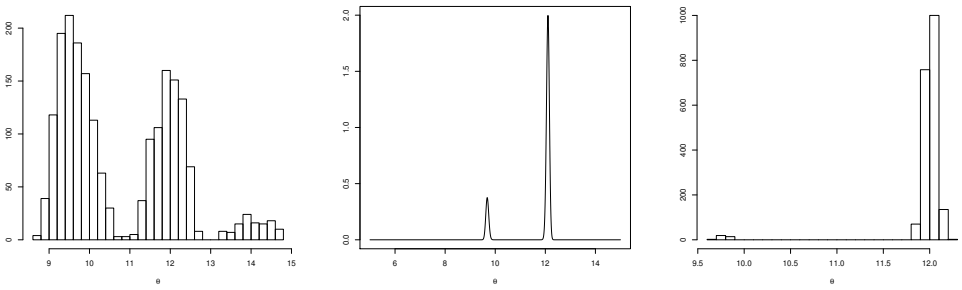
Unfortunately, sampling $\Pi(\boldsymbol{\theta}|\mathbf{z})$ using MCMC methods is unfeasible because the simulations are highly time-consuming. A good way for solving this problem is to switch $y_{\boldsymbol{\theta}}(\mathbf{x})$ with a Gaussian process in the likelihood expression (Cox et al., 2001; Kennedy and O'Hagan, 2001). Such an emulator is fitted from a learning set of simulations $\mathbf{y}(\mathbf{D_N})$ run in a design of experiments $\mathbf{D_N}$. The emulator then provides a cheap prediction of the model over the input space and MCMC algorithms can be performed by sampling from an approximated *posterior* distribution $\tilde{\Pi}$:

$$\tilde{\Pi}\big(\boldsymbol{\theta}|\mathbf{z}, y(\mathbf{D_N})\big) \propto \tilde{L}\big(\mathbf{z}|\boldsymbol{\theta}, y(\mathbf{D_N})\big)\Pi(\boldsymbol{\theta}). \tag{5}$$

When the Gaussian emulator yields good predictions of the model, the *posterior* distribution sampled from this approximated likelihood (5) is expected to be close to the true one (3). Owing to a poor emulator quality, however, results can be very disappointing. Indeed, such a calibration causes a new source of uncertainty which strongly depends on the design of experiments $\mathbf{D_N}$

used to fit the emulator. The default strategy consists in building $\mathbf{D_N}$ as a Space-Filling Design, for instance a maximin Latin Hypercube (Morris and Mitchell, 1995). In the case where the computer model is highly non linear, it is quite difficult to know in advance how many simulations are required to fit a good emulator. In this context, running sequentially the model has proven efficient for improving the predictive capability of the emulator or else for finding the optimum of the model. In the same spirit, we propose a new method for the calibration of costly computer models by building sequential designs. New algorithms are introduced with the help of the *EI* (*Expected Improvement*) criterion (Jones et al., 1998). It needs to be computed over the sum of square of the residuals $SS(\boldsymbol{\theta})$. Numerical illustrations in several dimensions are provided to assess the efficiency of such adaptive strategies in terms of reducing the Kullback Leibler discrepancy between the unknown Π and $\tilde{\Pi}$ (see Figure 1).

Figure 1: *The calibration of* $y_\theta(x) = (6x-2)^2 \times \sin(\theta x - 4)$, $\theta = 12$, $\lambda^2 = 0.09$. *Left: sampling of* $\tilde{\Pi}(\theta|z, y(D_{N=30}))$ *using a maximin LHD. Middle: the true posterior* $\Pi(\theta|z)$. *Right: sampling of* $\tilde{\Pi}(\theta|z, y(D_{N=30}))$ *using a sequential design based on the EI criterion.*



References

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, London, 1 edition, 1994.

K. Campbell. Statistical calibrations of computer simulations. *Reliability Engineering and System Safety*, 91:1358–1363, 2006.

D.D. Cox, J.S. Park, and E.S. Clifford. A statistical method for tuning a computer code to a data base. *Computational Statistics and Data Analysis*, 37:77–92, 2001.

D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

M. Kennedy and A. O'Hagan. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, 63:425–464, 2001.

D. Morris and J. Mitchell. Exploratory designs for computationnal experiments. *Journal of Statistical Planning and Inference*, 43:381–402, 1995.

Short biography – I have a master's degree in probability and statistics from the University Lille 1. My PhD takes place at EDF R&D in the Department of Industrial Risks Management. My contributions deal with the assessment of uncertainty in computer experiments. I propose some new approaches for both the calibration and the validation of costly computer models.

Optimization of inspection plans for structures submitted to stochastic degradation processes.

R. Décatoire
Phimeca Engineering - GeM Nantes University - I2M Bordeaux University

Supervisor(s): Prof. F. Schoefs (GeM Nantes University), Dr T. Yalamas (Phimeca Engineering) and Associate Prof. S.M. Elachachi (I2M Bordeaux University)

Ph.D. expected duration: 2012-2015

Adress: Phimeca Engineering 34 rue de Sarlieve 63800 Cournon d'Auvergne FRANCE

Email: decatoire@phimeca.com

Abstract:

The development of modern societies has seen the construction of numerous structures and infrastructures such as bridges, nuclear power plants or block of flats. Those structures, so-called "moderns" (i.e. built in the post-war years) are in reinforced concrete and subjected to important pathologies. Their management in order to ensure and guarantee their safety and durability is becoming a true economical challenge, in a context of limited budget: we thus speak of optimization of resource allocation under some requirements (security, ...) . It is therefore mandatory to optimize the inspection, maintenance and repair plans depending on the evolution of degradation indexes, in order to ensure the reliability of these structures [2, 1, 3].

From knowledge of the material characteristics of reinforced concrete, and their related uncertainties, it is possible to predict the evolution of different degradation processes such as concrete carbonation. Based on a degradation state discretization which gives a decision support in classes, sufficient for the stakeholders, the evolution of a state criterion is derived from physical degradation models and different degradation indexes are defined. Since we know the strong variability of reinforced concrete as well as the size of the concerned structure, this methodology takes into account the spatial variability of the material parameters. These predictions of the degradation aim at optimizing the inspection processes of these structures performed by non-destructive evaluation or testing (NDE or NDT), triggered from a reliability threshold of the structure. The number of measures and their location are determined with an adaptative design of experiments (ADoE) led by two criteria :

- The measures are sufficiently spaced,
- Based on the degradation models, the location which present the highest degradation is to be inspected in priority.

From this inspection plan which aim at evaluating the stochastic field of the degradation process at a given time, the probability of declaring that the structure is to be maintened can be evaluated with the help of the simulations made at the inspection locations. By adding the delay between inspections as an optimization variable, it is possible to optimize the operating cost of the concerned structures.

At the inspection date, the results coming from the NDT are likely to give direct information on the material parameters of the structure, or indirect by measuring the output of a degradation model which depends on those material parameters. These results are used in a Bayesian framework in order to update our knowledge of the structure. A Sobol' sensitivity analysis allows choosing which variables are to be updated depending on their importance, to ensure that the quality of degradation predictions is perceptibly increased. The measurement errors on the inputs and

output of the degradation model are taken into account. The inspection plan defined beforehand can thus be updated and the aDoE can take advantage of the measured outputs of a degradation model. This output are also used to classify the structure in two classes:

- location where the degradation is lower than a given threshold,
- location where the degradation is degradation higher than this given threshold,

with the help of support vector machine (SVM). The result of this classification will be the decision tool to be used by the stakeholders to decide if the considered structure needs to be repaired.

This method, illustrated in figure 1 will thus result in a full decision-making process allowing to rationalize the operating cost of their facilities.

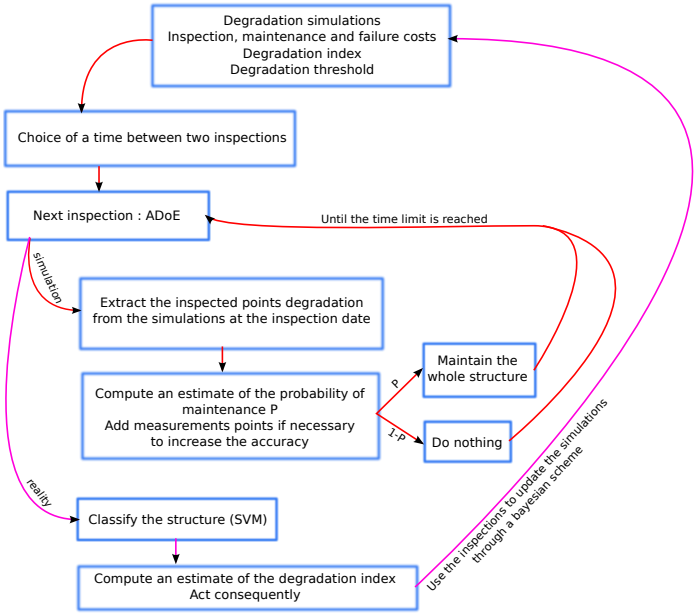


Figure 1: Illustration of the methodology

References

[1] D.M. Frangopol, A. Strauss, and K. Bergmeister. Lifetime cost optimization of structures by a combined condition-reliability approach. *Engineering Structures*, 1572-1580, 2009.

[2] M-J. Kallen. *Markov processes for maintenance optimization of civil infractructure in the Netherlands*. PhD thesis, Delft University of Technology, 2007.

[3] E. Sheils, A. O'Connor, D. Breysse, F. Schoefs, and S. Yotte. Development of a two-stage inspection process for the assessment of deteriorating infrastructure. *Reliability Engineering and System Safety*, 95:182–194, 2010.

Short biography – Engineer of the French Institut for Advanced Mechanics (IFMA) graduated in 2012. This CIFRE thesis is co-funded by Phimeca Engineering and the ANRT. It is linked with the French research project ANR EVADEOS which deals with the evaluation of concrete structures degradation and with their management. The aim of the thesis is to propose a method to optimize the inspection plans for such structures, taking into account the spatial variability of concrete properties.

A Bayesian approach to constrained multi-objective optimization of expensive-to-evaluate functions

P. Feliot
IRT SystemX / Supelec

Supervisor(s): J. Bect (Supelec), E. Vazquez (Supelec)

Ph.D. expected duration: 2014-2017

Adress: IRT SystemX, 8 avenue de la Vauve, 91120 Palaiseau

Email: paul.feliot@irt-systemx.fr

Abstract:

We address the problem of derivative-free multi-objective optimization of real-valued functions under multiple inequality constraints:

$$\begin{cases} \text{Minimize} & f(x) \\ \text{Subject to} & x \in \mathbb{X} \quad \text{and} \quad c(x) \leq 0 \end{cases}$$

where $f = (f_j)_{1 \leq j \leq p}$ is a vector of objective functions to be minimized, $\mathbb{X} \subset \mathbb{R}^d$ is the search domain and $c = (c_i)_{1 \leq i \leq q}$ is a vector of constraint functions. Both the objective functions f_j and the constraint functions c_i are assumed to be smooth, nonlinear functions that are expensive to evaluate. As a consequence, the number of evaluations that can be used to carry out the optimization is very limited. This setup typically arises when the values $f(x)$ and $c(x)$ for a given $x \in \mathbb{X}$ correspond to the outputs of a computationally expensive computer program.

In this work, we consider a Bayesian approach to this optimization problem. The objective and constraint functions are modelled using a vector-valued Gaussian process and $\mathbb{X}$ is explored using a sequential Bayesian design of experiments approach. More specifically, we focus on the Expected Improvement (EI) infill sampling criterion. This criterion was originally introduced in the context of single-objective, unconstrained optimization [5]. It was later extended to handle constraints [8, 9] and to address unconstrained multi-objective problems [7, 11, 4]. However, to the best of our knowledge, the general case of a constrained multi-objective problem has only been addressed very recently by [10]. In their paper, Shimoyama et al. consider three different Bayesian criteria for unconstrained multi-objective optimization and study the effect of multiplying the criteria by a probability of feasibility in order to handle the constraints.

The approach we propose to handle the constraints is based on an extended domination rule, in the spirit of [2, 6], which takes both objectives and constraints into account under a unified framework. The extended domination rule makes it possible to derive a new hypervolume based expected improvement criterion to deal with constrained multi-objective optimization problems. The new criterion is equivalent to the original EI on unconstrained single-objective problems and to Schonlau's extension to the constrained case [9] once a feasible point has been found. It is also similar to the formulation of [11] for unconstrained multi-objective problems and to that of [10] in the constrained case once a feasible point has been found. As such, it can be seen as a generalization of the above-mentioned criteria.

The calculation of this class of criteria is known to become difficult as the number of objectives increases [3]. To address this difficulty, we propose a novel approach, making use of Sequential Monte Carlo techniques. Moreover we also use Sequential Monte Carlo techniques for the optimization of our new criterion [1]. The performance of the proposed method is evaluated on a set of test problems coming from the literature and compared with reference methods.

References

[1] R. Benassi, J. Bect, and E. Vazquez. Bayesian optimization using sequential Monte Carlo. In *Learning and Intelligent Optimization. 6th International Conference, LION 6, Paris, France, January 16-20, 2012, Revised Selected Papers*, volume 7219 of *Lecture Notes in Computer Science*, pages 339–342. Springer, 2012.

[2] Carlos M Fonseca and Peter J Fleming. Multiobjective optimization and multiple constraint handling with evolutionary algorithms. I. A unified formulation. *IEEE Transactions on Systems, Man and Cybernetics. Part A: Systems and Humans*, 28(1):26–37, 1998.

[3] Iris Hupkens, Michael Emmerich, and André Deutz. Faster computation of expected hypervolume improvement. *arXiv preprint arXiv:1408.7114*, 2014.

[4] Shinkyu Jeong, Youichi Minemura, and Shigeru Obayashi. Optimization of combustion chamber for diesel engine using kriging model. *Journal of Fluid Science and Technology*, 1(2):138–146, 2006.

[5] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[6] Akira Oyama, Koji Shimoyama, and Kozo Fujii. New constraint-handling method for multi-objective and multi-constraint evolutionary optimization. *Transactions of the Japan Society for Aeronautical and Space Sciences*, 50(167):56–62, 2007.

[7] V. Picheny. Multiobjective optimization using Gaussian process emulators via stepwise uncertainty reduction. *Statistics and Computing*, DOI:10.1007/s11222-014-9477-x:1–16, 2014.

[8] V. Picheny. A stepwise uncertainty reduction approach to constrained global optimization. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS), 2014, Reykjavik, Iceland.*, volume 33, pages 787–795. JMLR: W&CP, 2014.

[9] M. Schonlau, W. J. Welch, and D. R. Jones. Global versus local search in constrained optimization of computer models. In *New Developments and Applications in Experimental Design: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference*, volume 34 of *IMS Lecture Notes-Monographs Series*, pages 11–25. Institute of Mathematical Statistics, 1998.

[10] Koji Shimoyama, Koma Sato, Shinkyu Jeong, and Shigeru Obayashi. Updating kriging surrogate models based on the hypervolume indicator in multi-objective optimization. *Journal of Mechanical Design*, 135(9):094503, 2013.

[11] T. Wagner, M. Emmerich, A. Deutz, and W. Ponweiser. On expected-improvement criteria for model-based multi-objective optimization. In *Parallel Problem Solving from Nature, PPSN XI. 11th International Conference, Krakov, Poland, September 11-15, 2010, Proceedings, Part I*, volume 6238 of *Lecture Notes in Computer Science*, pages 718–727. Springer, 2010.

Acknowledgements. This research work has been carried out in the frame of the Technological Research Institute SystemX, and therefore granted with public funds within the scope of the French Program *Investissements d'Avenir*.

Short biography – I am titular of an engineering diploma from the french engineering school ISAE-Supaero (Toulouse) and of a Master of Research from the Paul Sabatier University (Toulouse). I started my PhD thesis with Supelec at the Technological Research Institute SystemX in January 2014 under the ROM project (model Reduction and Multi-physic Optimization). Industrial partners of the project are SAFRAN (Snecma), Cenaero, Airbus group, Renault and ESI group.

Sensitivity indices for the exploration of the input domain

J. Fruth
TU Dortmund University

Supervisor(s): Prof. S. Kuhnt (Dortmund University of Applied Sciences and Arts) and Prof. O. Roustant (Ecole des Mines, Saint Etienne)

Ph.D. expected duration: 2012-2015

Adress: Faculty of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany

Email: fruth@statistik.tu-dortmund.de

Abstract: We discuss two new sensitivity indices, the *first-order support index* and the *total support index*, which give a solution to the question formulated in [3] of "how to determine regions within in the input space for which the model variation is high". The indices are defined as functions over the domain of input variables and give insight into the local influence of the variables over the whole domain. The result is global in the sense that the other variables are not fixed but varied over the global space. The method provides an informative extension to a standard sensitivity analysis and can in addition be especially helpful in the specification of the input domain, a critical, but often vaguely handled issue in sensitivity analysis.

As an example, consider the well known Ishigami function by [2],

$$f(x_1, x_2, x_3) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4 \sin(x_1).$$

Table 1 shows different settings of the input distributions (μ_1, μ_2, μ_3) together with the corresponding unscaled total sensitivity indices (D_1^T, D_2^T, D_3^T). The first row shows the usual setting, resulting in X_1 having the highest index followed by X_2 and X_3. In the second row, we reduced the support of the distribution of X_3 by 10 percent. Now, X_1 and X_3 have rapidly lost influence in comparison to X_2 and the ranking of variables has changed. This extreme difference can be explained by the interaction between X_1 and X_3, which is much stronger at the borders than in the rest of the function while the behaviour of X_2 stays unchanged.

μ_1	μ_2	μ_3	D_1^T	D_2^T	D_3^T
$U(-\pi, \pi)$	$U(-\pi, \pi)$	$U(-\pi, \pi)$	7.72	6.13	3.37
$U(-\pi, \pi)$	$U(-\pi, \pi)$	$U(-\pi+\frac{\pi}{10}, \pi-\frac{\pi}{10})$	4.05	6.13	1.45

Table 1: Unscaled total sensitivity indices of the Ishigami function for different input distributions.

This behavior can be analysed and visualized by support index functions [1], which examine the influence of the input variables at each point of the input domain t. We assume a situation $Y = f(X_1, \ldots, X_d)$ with X_i independent continuous random variables and $f : [0,1]^d \to \mathbb{R}$ in C^1.

Definition. *Support index functions*
The first-order support index $D_i(t)$ of an input variable X_i at a point t, $t \in [0,1]$, is defined as the square of the expected value of the first derivative of f,

$$D_i(t) = \left(E\left(\frac{\partial f}{\partial \mathbf{x}_i}(t, \mathbf{X}_{-i}) \right) \right)^2.$$

The total support index $D_i^T(t)$ of an input variable X_i at a point t, $t \in [0,1]$, is defined as the expected value of the squared derivative of f,

$$D_i^T(t) = E\left(\left(\frac{\partial f}{\partial \mathbf{x}_i}(t, \mathbf{X}_{-i})\right)^2\right).$$

In the general framework of independent continuous input variables, we present theoretical results about the asymptotical connection between the support indices and their scalar equivalents. Precisely, we show that the first-order support index of a variable X_i is linked to the first-order Sobol index of X_i, when we truncate the distribution of X_i to $[t - \frac{h}{2}, t + \frac{h}{2}]$ and let h tend to zero. A similar link exists between the total support index and the total sensitivity index. We also give an interpretation for the area below the functions, i.e. the expected value of the functions over the input domain. Finally, we present a succesful application in the field of sheet metal forming.

References

[1] J. Fruth, O. Roustant, and S. Kuhnt. Support indices: Measuring the effect of input variables over their support, 2015. Online: https://hal.archives-ouvertes.fr/hal-01113555.

[2] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computer models. In B. M. Ayyub, editor, Proceedings of the ISUMA '90, pages 398–403, 1990.

[3] A. Saltelli, K. Chan, and E.M. Scott. Sensitivity analysis. Wiley, Chichester, 2000.

Short biography – Jana Fruth did her Master of Statistics at TU Dortmund University in Dortmund, Germany. In her PhD thesis she works on sensitivity analysis methods for various situations in sheet metal forming, including interaction analysis, analysis of functional inputs, and derivative-based indices. The project is a cooperation with the TU Dortmund Engineering Department as part of the collaborative research centre SFB 708, funded by the Deutsche Forschungsgemeinschaft (DFG).

Sobol' indices estimation using nested designs

L. Gilquin
University of Grenoble

Supervisor(s): C. Prieur (University of Grenoble), E. Arnaud (University of Grenoble)

Ph.D. expected duration: 2013-2016

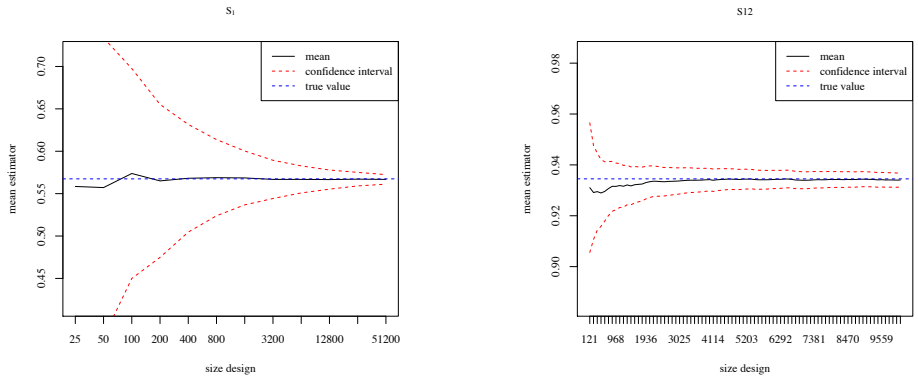
Adress: 655 Avenue de l'Europe, 38330 Montbonnot-Saint-Martin

Email: laurent.gilquin@inria.fr

Abstract: In many mathematical models, input parameters often poorly-known can have significant effects on the output of the model. It is important for model users to measure these effects. Global sensitivity analysis is a common practice to identify influent inputs and detect the potential interactions between them. Among the large number of available approaches, the variance-based method introduced by Sobol' (1993) allows to calculate sensitivity indices called Sobol' indices. In the same paper Sobol' proposes a "pick-freeze" estimator of the indices. This estimation procedure has a linear cost in the input space dimension to estimate all first-order indices. This linear dependency disappears when using replicated designs as the number of runs becomes independent of the input space dimension. A synthesis on the use of replicated designs (referred as permuted column sampling plans) can be found in Morris *et al.* [4] where the authors use the approach introduced by McKay (1995). Later on, Mara *et al.* [3] combine replicated designs with "pick-freeze" estimators to estimate first-order Sobol' indices. This procedure has been further studied (asymptotic properties for first-order indices) and generalized in Tissot *et al.* [6] to the estimation of closed second-order indices. For closed second-order indices, the procedure relies on the replication of randomized orthogonal arrays.
Since the budget of runs of model users is often restricted by a threshold, it is important to assess the minimal number of runs required to obtain proper indices estimations. In case where the number of points in the initial design is not sufficient to give proper results, an iterative process can be used. The initial design is iteratively augmented such that the new responses are added to the former ones. This iterative process can be simply applied in the case of classical procedures (Sobol', Saltelli) contrary to the replication method where no solutions has not yet been proposed in the literature.

We propose here to render the replication method iterative. We develop an iterative process using a nested design (also referred as augmented design) to estimate, with the replication method, first-order or closed-second order Sobol' indices. The nested design is specific to each case. The iterative process consists in augmenting the nested design until the precision on the estimated indices satisfies a stopping criterion such as a threshold on the absolute value of the discrepancy between two consecutive estimations.
For the estimation of first-order Sobol' indices, we can exploit the nested design introduced in Qian [5]. This design consists of a nested Latin hypercube with multiple layers. This special Latin hypercube design contains a succession of embedded Latin hypercubes.
The construction of a nested design for the estimation of closed second-order Sobol' indices relies on the construction of nested orthogonal arrays of strength two (2-NOA). The construction of an orthogonal array of strength two (2-OA) is easily achieved when the number of levels of the 2-OA is a prime number. Unfortunately, given q_1 and q_2 two different prime numbers the corresponding 2-OAs cannot be embedded. Thus, the method proposed by Qian cannot simply be extended to construct a 2-NOA. Alternatives have been studied but the proposed nested designs either possess a restrictive threshold for the number of parameters or do not discretize each parameter equally (asymmetric orthogonal array).

We propose two methods to construct a 2-NOA with the aim of estimating closed second order Sobol' indices through the previously described iterative process. The first method is stochastic and relies on results from graph theory [1]. The second method consists in duplicating an initial 2-OA. Then, for each copy of the 2-OA, the levels are relabeled with a different set of values.
The standard method (without a nested design) to estimate closed-second order Sobol' indices is to construct a 2-OA with a large number of levels. We compare the space-filling properties of the two nested designs obtained from each method to the 2-OA used in the standard method for an equivalent number of runs. The comparison is based on different criteria such as discrepeancy measure, Eglai's criterion, mindist, Kullback-Leibler divergence [2] and euclidean minimal spanning tree. Then, we conduct numerical experimentations on classical test functions to compare the closed second-order Sobol' indices obtained with our two methods to the one obtained with the standard method. The figure below shows results for first and closed second-order Sobol' indices, in the case of the g-Sobol' function, estimated with the first method, with 95% confidence interval.



(a) first-order index

(b) closed second-order index

References

[1] M. Depolli, J. Konc, K. Rozman, R. Trobec, and D. Janežič. Exact parallel maximum clique algorithm for general and protein graphs. *J. Chem. Inf. Model.*, 53:2217–2228, 2013.

[2] A. Jourdan and J. Franco. A new criterion based on kullback-leibler information for space filling designs. *https://hal.archives-ouvertes.fr/hal-00375820/*, 2009.

[3] T. A. Mara and O. Rakoto-Joseph. Comparison of some efficient methods to evaluate the main effect of computer model factors. *J. Statist. Comput. Simulation*, 78:167–178, 2008.

[4] M. Morris, L. M. Moore, and M. D. McKay. Orthogonal arrays in the sensitivity analysis of computer models. *Technometrics*, 50:205–215, 2008.

[5] P. Z. G. Qian. Nested latin hypercube designs. *Biometrika*, 96:957–970, 2009.

[6] J. Y. Tissot and C. Prieur. A randomized orthogonal array-based procedure for the estimation of first- and second-order sobol' indices. *To appear in Journal of Statistical Computation and Simulation*, 2014.

Short biography – Laurent Gilquin is a PhD student in Applied Mathematics from University Grenoble Alpes. The context of his thesis is the analysis through stochastic tools of the LUTI (Land Use and Transport Integrated) model TRANUS. This work is supported by the CITiES project funded by the Agence Nationale de la Recherche (grant ANR-12-MONU-0020). The CITiES project has for topic the calibration and the validation of LUTI models.

Robust construction of a spatio-temporal surrogate model - Application in thermal engineering

J. Guerra
ONERA - Epsilon Ingénierie

Supervisor(s): F. Gamboa (Institut de Mathématiques de Toulouse), P. Klotz (ONERA)

Ph.D. expected duration: 2013-2016

Adress: 2, avenue Edouard Belin - 31055 Toulouse Cedex 4

Email: jonathan.guerra@onera.fr

Abstract: The objective is to develop a robust construction of a spatio-temporal surrogate model capable of long term in time predictions, in order to replace an accurate but time expensive numerical simulation model of a thermal transient phenomenon. Some industrial applications can require an extensive use of numerical simulations: for instance optimization processes or inverse problems are rather greedy. In addition, the numerical simulations at disposal are often too much CPU-time demanding to be used in those cases.

Subsequently, we have to propose a robust construction of a spatio-temporal surrogate model which is able to predict a transient thermal phenomenon, i.e. a first-order in time one. To do that, the statistical learning theory is particularly appropriate. In this case, a surrogate model is an analytical model low cost to evaluate, which is built by minimizing the mean squared error it commits on a set of observations. In the spatio-temporal case, the observations are time-discretized trajectories of the inputs-outputs couple generated with the costly reference model. If those trajectories are chosen smartly, it is possible to construct a model able to predict transient response to various boundary conditions. Because the reference model can be complex and costly, the construction introduced here has to deal with several constraints: firstly, the number of learning trajectories is small, secondly the input dimension of the surrogate can be large, thirdly it has to execute long-term in time predictions.

Let's start from the physical problem: we seek to predict the behaviour of a function $y(\boldsymbol{p}, t)$ (with $\boldsymbol{p} \in \mathbb{R}^3$ the spatial position and $t \in \mathbb{R}$ the time variable) obeing to a first-order phenomenon:

$$\frac{\partial y(\boldsymbol{p}, t)}{\partial t} = \mathring{\boldsymbol{F}}\Big(y(\boldsymbol{p}, t), \boldsymbol{u}(t)\Big) \tag{1}$$

With $\boldsymbol{u}(t)$ the functional vector of N_u boundary conditions and forcing terms. For instance in thermal engineering, $y(\boldsymbol{p}, t)$ can represent the temperature in an electronic equipment and $\boldsymbol{u}(t)$ the power to dissipate or the ambiant temperature, (1) being the heat equation.

Let's now discretize spatially and temporally this equation. Spatially, let $(\boldsymbol{p}_1, \cdots, \boldsymbol{p}_{N_y})$ be the position of N_y points of interest. Then, $\forall j \in [\![1, N_y]\!]$ $y_j(t) := y(\boldsymbol{p_j}, t)$ and $\boldsymbol{y}(t) = \big(y_1(t), \cdots, y_{N_y}(t)\big) \in \mathbb{R}^{N_y}$. Temporally, let's consider the following discretization of $[0, T]$: $t^k = k\Delta t$ (time step is considered constant) and $t^{N_t} = T$. Thence, $\boldsymbol{y}^k := \boldsymbol{y}(t^k)$ and $\boldsymbol{u}^k := \boldsymbol{u}(t^k)$. Thanks to those notations, an explicit discretization of (1) implies:

$$\boldsymbol{y}^k = \boldsymbol{F}(\boldsymbol{y}^{k-1}, \boldsymbol{u}^k) \tag{2}$$

From this, the mathematical form of the surrogate model $\hat{\boldsymbol{F}}$ of $\boldsymbol{F}$, parametrized by $\boldsymbol{w} \in \mathbb{R}^{N_w}$, can be introduced as follows:

$$\begin{cases} \hat{\boldsymbol{y}}^k = \hat{\boldsymbol{F}}\left(\hat{\boldsymbol{y}}^{k-1}, \boldsymbol{u}^k; \boldsymbol{w}\right) \\ \hat{\boldsymbol{y}}^0 = \boldsymbol{y}^0 \end{cases} \tag{3}$$

With $\boldsymbol{y}^0 := (y_1(0), \cdots, y_{N_y}(0))$ the initial condition of the transient phenomenon (given by the user).

It is important to notice here that the framework of dynamic neural network [3] is particularly adapted to the equation (3). Thus, let's consider a one-hidden-layer perceptron for $\hat{\boldsymbol{F}}$. Let $\boldsymbol{s} \in \mathbb{R}^{N_y}$ be the output of the neural network $\hat{\boldsymbol{F}}$: $(\boldsymbol{s})_{j \in [\![1, N_y]\!]} = \left\{ \hat{\boldsymbol{F}}(\boldsymbol{x}\,; \boldsymbol{w}) \right\}_{j \in [\![1, N_y]\!]}$. Its mathematical definition is:

$$s_j = \sum_{n=1}^{N_n} w_{0nj} \phi \left(\sum_{i=1}^{N_u+N_y} w_{in0} x_i + w_{0n0} \right) + w_{00j}$$

With $\boldsymbol{x} \in \mathbb{R}^{N_u+N_y}$ the input vector of the neural network, N_n the number of neurons in the hidden layer (the complexity of a perceptron), ϕ is the sigmoidal function defined $\forall z \in \mathbb{R}$ by $\phi(z) = \frac{2}{1+e^{-2z}} - 1$ and $\boldsymbol{w} = (w_{inj})_{\substack{0 \leqslant i \leqslant N_u+N_y \\ 0 \leqslant n \leqslant N_n \\ 0 \leqslant j \leqslant N_y}}$, the weights of the perceptron. $\boldsymbol{w} \in \mathbb{R}^{N_w}$ with $N_w = (N_u + N_y + 1) N_n + (N_n + 1) N_y$. In other words, the surrogate model $\hat{\boldsymbol{F}}$ can predict the state of the system at time t^k thanks to the information of the state of the system at time t^{k-1}. This system being controlled by the value of the exogenous variables $\boldsymbol{u}^k$. To compute the weights, the following minimization has to be performed:

$$\min_{\boldsymbol{w} \in \mathbb{R}^{N_w}} \underbrace{\sum_{l=1}^{N_l}}_{\text{trajectory}} \underbrace{\sum_{k=1}^{N_t}}_{\text{timestep}} \left\| \boldsymbol{y}^{k,l} - \hat{\boldsymbol{F}} \left(\hat{\boldsymbol{y}}^{k-1,l}, \boldsymbol{u}^{k,l}\,; \boldsymbol{w} \right) \right\|_2^2 \tag{4}$$

Now, this recurrent neural network methodology has to be improved to answer to the industrial constraints:

- The optimization of the weights can be difficult because of the dimension of the problem. A multilevel optimization is then used to accelerate the process. Instead of solving the classical problem in (4), a decomposition output by output splits the large problem to smaller and easier ones which are solved iteratively. And because this resolution is fully distributed, it also allows to build the network faster.
- To overcome the limited number of learning trajectories available, a cross validation technique [1] is used for model selection. To be adapted to the spatio-temporal case, an innovative way to split the observations in the folds is proposed.
- Other solution to this small budget of trajectories, a spatio-temporal design of experiments is introduced, by generalizing a static criterion [2] in the dynamic case.
- Sensitivity analysis theory [4] is also applied in this case to reduce the input dimension.

References

[1] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

[2] Kai-Tai Fang, Runze Li, and Agus Sudjianto. *Design and modeling for computer experiments*. CRC Press, 2010.

[3] Matthias De Lozzo, Patricia Klotz, and Béatrice Laurent. Multilayer perceptron for the learning of spatio-temporal dynamics - application in thermal engineering. *Engineering Applications of Artificial Intelligence*, 26(10):2270 – 2286, 2013.

[4] Andrea Saltelli, Karen Chan, E Marian Scott, et al. *Sensitivity analysis*, volume 134. Wiley New York, 2000.

Short biography – I was graduated from a master degree on Partial Differential Equations, modeling and scientific computing and from a school of engineering with a specialization in aerodynamics and numerical simulation. My PhD is now focusing on optimization under uncertainty of transient thermal phenomena.

Kriging based sequential optimization with mixed qualitative and quantitative inputs

Momchil Ivanov
TU Dortmund University

Supervisor(s): Prof. Sonja Kuhnt (University of Applied Sciences and Arts, Dortmund) and Prof. Christoph Buchheim (TU Dortmund University)

Ph.D. expected duration: 2011-2015

Adress: Department of Computer Science, University of Applied Sciences and Arts, Emil-Figge-Str. 42, 44227 Dortmund

Email: ivanov@statistik.tu-dortmund.de

Abstract: In many applications in industry it is nowadays turning into standard practice to study complex processes with the help of computer experiments. With increasing computing capabilities it has become customary to perform simulation studies beforehand, where the desired process characteristics can be optimized. However, simulations generally take a long time to run, ranging from hours to days, making it impossible to perform direct optimization on the computer code. Instead, the simulator can be considered as a black-box function and a (meta-)model, which is cheaper to evaluate, is used to interpolate the simulation.

The optimization of expensive to evaluate black-box functions is often performed with the help of model-based sequential strategies. A popular choice is the efficient global optimization (EGO) algorithm, which is based on the prominent Kriging metamodel [4]. Kriging allows a great flexibility and can be used to approximate highly non-linear functions. It also provides a local uncertainty estimator at unknown locations, which can be used to guide the EGO algorithm to less explored regions of the search space. EGO based strategies have been applied in numerous simulation studies with great success, however, a big drawback of Kriging and, by extension, the classical EGO algorithm is that the existing method is only able to deal with continuous input data.

It should come as no surprise, that there are numerous application examples from research or industrial fields which have mixed continuous/ordered/qualitative inputs. One example is a forwarding-facility computer experiment described in [5]. The qualitative inputs could be represented by distinct vehicle loading or unloading strategies which logically lead to different final outputs. Naturally in this mixed discrete-continuous situation we would still like to use the Kriging method and take advantage of the many valuable qualities it has, like the uncertainty estimates it provides. Very few publications have up to now scratched the topic of mixed discrete-continuous modeling and optimization, e.g., Zhou et al. [7] invent a special Kriging kernel for the mixed case and Swiler et al. [6] continue their work. However, there seem to be even less contributions for Kriging-based sequential optimization.

In this talk we present a novel way to transform the Kriging kernel, based on the mixed discrete-continuous Gower distance [2]. The Gower distance represents a simple, but effective concept of defining a distance measure between qualitative and quantitative objects. It also has appealing semi-definiteness properties, making it a natural choice of a distance measure to use in covariance estimation. The proposed kernel modification is very flexible and allows an easy adaptation to the mixed inputs case of any of the classical Kriging kernels. Definition 1 shows the adjustments made to the simple, but often used exponential covariance kernel. This modified Gower-exponential kernel allows for Kriging modeling in the discrete case, while retaining its anisotropic property (the scales θ_i are estimated for both the continuous and the categorical dimensions). In our approach the number of kernel parameters to estimate scales only linearly with the number qualitative

factors, as opposed to the exponential growth in the work of Zhou et al. [7]. This makes our method computationally tractable for applications with more than a small number of categorical inputs. It also allows sensitivity analysis methods, like the FANOVA decomposition to be applied, enabling us to estimate the interaction structure between the categorical and the numerical variables. This makes the dimensionality reduction and parallel optimization approach introduced by Ivanov and Kuhnt [3] feasible in the mixed case. Furthermore, as the estimation of the kernel parameters directly implies a data-dependent Gower distance function for the mixed space, the parallelization methods proposed by Bischl et al. [1] become applicable.

Definition 1 ***(Gower-exponential covariance function):***
Let $\mathcal{F} := \mathcal{D} \times \mathcal{Z}$ be a mixed k-dimensional space, where $\mathcal{D} \subseteq \mathbb{R}^n$ and $\mathcal{Z}$ is an m-dimensional space containing the discrete/categorical variables. For two data points $\mathbf{x}, \mathbf{x}' \in \mathcal{F}$ the Gower-exponential kernel is defined as:

$$Cov(Z(\mathbf{x}), Z(\mathbf{x}')) = \exp\left(\sum_{i=1}^{k} -\theta_i d_i\left(\mathbf{x}, \mathbf{x}'\right)\right)$$

where d_i represents i-th component function of the Gower distance.

In this presentation we also show a new, computationally cheap heuristic strategy to maximize the expected improvement infill criterion (an important building block of EGO) in the mixed discrete-continuous case. The usefulness of this novel mixed EGO algorithm is shown in this talk with the help of a few benchmark functions. A discussion of the further challenges and research directions in this exciting and rather underrated field of computer experiments with mixed qualitative and quantitative inputs concludes the presentation.

References

[1] B Bischl, S Wessing, N Bauer, K Friedrichs, and C Weihs. MOI-MBO: Multiobjective infill for parallel model-based optimization. In *Learning and Intelligent Optimization*, Lecture Notes in Computer Science, pages 173–186. Springer International Publishing, 2014.

[2] JC Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

[3] M Ivanov and S Kuhnt. A parallel optimization algorithm based on fanova decomposition. *Quality and Reliability Engineering International*, 30(7):961–974, 2014.

[4] DR Jones, M Schonlau, and WJ Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[5] L Neumann and S Deymann. Transsim-node - a simulation tool for logistics nodes. In *Proceedings of the Industrial Simulation Conference 2008*, pages 283–287, 2008.

[6] LP Swiler, PD Hough, P Qian, X Xu, C Storlie, and H Lee. Surrogate models for mixed discrete-continuous variables. In *Constraint Programming and Decision Making*, pages 181–202. Springer, 2014.

[7] Q Zhou, PZG Qian, and S Zhou. A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics*, 53(3):266–273, 2011.

Short biography – I have studied mathematics at the Universitiy of Kaiserslautern (Germany). Since 2011 I am a PhD student at the TU Dortmund university. There I began to work in the SFB 708 project with continuous industrial processes as research subject. In 2014 the focus of my research shifted towards mixed discrete-continuous simulations as I began to work in the RTG 1855 project. The topic of my work is meta-modeling and sequential black-box optimization.

Functional error modeling for Bayesian inference in hydrogeology

L. Josset
Institute of Earth Sciences, University of Lausanne

Supervisor(s): Prof. Ivan Lunati (University of Lausanne)

Ph.D. expected duration: 2011-2015

Adress: Institute of Earth Sciences, Geopolis - UNIL Mouline, CH-1015 Lausanne

Email: laureline.josset@unil.ch

Abstract:

The main challenge in groundwater problems stems from the lack of information about the underground properties. Indeed, aquifer conductivity and porosity fields are only known at few discrete locations, which makes impossible for hydrogeological problems to be addressed in a deterministic sense. What is generally done is to opt for a stochastic description of the underground and use Monte Carlo approaches [1] to propagate the uncertainty to the quantity of interest (for instance the breakthrough curve of a contaminant in a drinking well).

However, the propagation of the uncertainty requires to perform flow simulations for each of the underground realizations explored within the Monte Carlo set-up. When the number of realizations is high, the computational cost associated to the runs of the flow solver becomes prohibitive. State-of-the-art techniques often resort to approximate (or proxy) models, which allow to obtain an approximate flow response at considerably cheaper costs (by using upscaling techniques [2] or simplifying the description of the physical processes [8]). However, because of the numerous approximations, any inference based on the proxy model responses lead to biased conclusions and it is thus necessary to develop an error model to ensure reliable predictions.

We introduce a novel methodology to construct the error model by using Functional Data Analysis[7] in a Machine Learning approach. The error model consists in finding a mapping from the proxy to the exact responses. Its construction can be decomposed in three parts: first, the learning set of realizations is identified, and the corresponding approximate and exact responses are obtained. Second, Functional Principal Component Analysis (FPCA [4]) is used to decrease the dimensionality of the problem and help evaluate the suitability of the proxy model for the quantities of interest at hand. Third, the relationship between the two sets of curves is sought. To this end, we define a linear multivariate multiple regression model between the curves projections on the FPCA basis. Once the regression model is estimated on the learning set, it can be used to predict the exact response of any new geostatistical realization at limited computational cost. Indeed, the mapping gives us a prediction of the exact response using solely the approximate response.

The proposed methodology presents important computational advantages in several contexts. Considering the case of a non-aqueous phase liquid contamination problem (such as hydrocarbons spills), the evolution of the contamination plume is governed by a set of nonlinear transport equations leading to prohibitive computational costs. Using a single-phase transport problem as proxy in place of the computationally heavy two-phase solution and constructing an error model on a training set of 20 realizations, we are able to propagate the uncertainty described by the ensemble of 1000 curves and provide a reliable prediction of the quantiles at a fraction of the cost [6]. Beyond the scope of uncertainty propagation, the methodology is equally relevant in a Bayesian Inference context. For instance, in a two-stage MCMC [3] set-up, the proxy response corrected by the error model is used to decide whether or not it is worth to run the exact solver. As a result, the acceptance rate increases and we are thus able to run lengthier chains [5]. Nested

Sampling [9], an alternative to MCMC, is also considered, where resampling is performed at the prior level. Again, the error model is useful to reject sampled points. In addition, this approach allows the iterative update of the regression model as new flow simulations are performed. The performance of the error model is evaluated using the iterative construction in order to decide when it is sufficiently accurate and optimize the error model.

References

[1] G Dagan. An overview of stochastic modeling of groundwater flow and transport: From theory to applications. *Eos, Transactions American Geophysical Union*, 83(53):621, 2002.

[2] L J Durlofsky. Upscaling and gridding of fine scale geological models for flow simulation. In *8th International Forum on Reservoir Simulation, Italy*, pages 20–24, 2005.

[3] Y Efendiev, A Datta-Gupta, X Ma, and B Mallick. Efficient sampling techniques for uncertainty quantification in history matching using nonlinear error models and ensemble level upscaling techniques. *Water Resources Research*, 45(11), 2009.

[4] B Henderson. Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics*, 17(1):65–80, 2006.

[5] L Josset, V Demyanov, AH Elsheikh, and I Lunati. Accelerating monte carlo markov chains with proxy and error models. *under revision*, 2015.

[6] L Josset, D Ginsbourger, and I Lunati. Functional error modeling for uncertainty quantification in hydrogeology. *under revision*, 2015.

[7] J O Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

[8] C Scheidt and J Caers. Representing spatial uncertainty using distances and kernels. *Mathematical Geosciences*, 41(4):397–419, 2009.

[9] J Skilling. Nested sampling. *Bayesian inference and maximum entropy methods in science and engineering*, 735:395–405, 2004.

Short biography – After graduating from the EPFL with a master degree in physics, I joined the Ensemble project (ensemble-modeling.org) as a PhD student under the supervision of Ivan Lunati at the University of Lausanne. The project combined the expertise of sedimentologist, geostatisticians and geophysicists to tackle the question of stochastic ensemble aquifer modeling. More specifically, I have focused on the use of proxy and error models to simulate flow in porous media. The Ensemble Project is a synergia project funded by the Swiss FNS.

Sequential design of experiments oriented toward the quantile estimation using polynomial chaos expansions - application to the numerical dosimetry

Pierric Kersaudy
Université Paris-Est

Supervisor(s): Prof. Odile Picon (Université Paris-Est) and Dr. Joe Wiart (Orange Labs, Whistlab)

Ph.D. expected duration: 2012-2015

Adress: Orange Labs, Whistlab, 38 rue du Général Leclerc, 92130 Issy-les-Moulineaux

Email: pierric.kersaudy@orange.com

Abstract: If recent advances high-performance computation led to strongly reduce computational time for numerical dosimetry, the calculation of the Specific Absorption Rate that assess the human exposure to electromagnetic fields (EMF) remains time-consuming (a few hours per calculation). Consequently usual methods as Monte Carlo simulations cannot be used to analyze the influence of random input parameters variability on the SAR (the output variable) and the quantiles of the output distributions. Then, optimal meta-modeling strategies have to be employed to model the output response depending on the input parameters.

Let us consider the vector $\boldsymbol{x} = \{x_1, ..., x_m\} \in \mathbb{R}^M$ of M independent input parameters. In the polynomial chaos theory the uncertainty affecting $\boldsymbol{x}$ leads to its representation by a random vector $\boldsymbol{X}$ with prescribed probability density function $p_{\boldsymbol{X}}(\boldsymbol{x})$ in a probability space with probability measure $\mathbb{P}_{\boldsymbol{X}}$. The propagation of the uncertainty of $\boldsymbol{X}$ through a physical model M computationally expensive yields the random output variable $Y = M(\boldsymbol{X})$. Assuming that Y have finite variance and belongs to the Hilbert space of $\mathbb{P}_{\boldsymbol{X}}$-square integrable functionals of $\boldsymbol{X}$ with respect to the inner product $< \psi(\boldsymbol{X}), \phi(\boldsymbol{X}) >= \mathbb{E}(\psi(\boldsymbol{X})\phi(\boldsymbol{X}))$, Y can be expanded on an orthogonal polynomial basis which give the following *polynomial chaos expansion* [1].

$$Y = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^m} \beta_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \tag{1}$$

Where $\boldsymbol{\alpha} = [\alpha_1 ... \alpha_m]$ is the multi-index, the $\beta_{\boldsymbol{\alpha}}$ are deterministic coefficients to compute and the $\psi_{\boldsymbol{\alpha}}$ are multivariate orthogonal polynomials with respect to the above inner product. In this communication, the non intrusive regression method is used to estimate the coefficients $\beta_{\boldsymbol{\alpha}}$ of a sparse representation of the expansion. The *least-angle regression* algorithm (LARS) is used to select a sparse set of influential polynomials [1] and the *leave-one-out cross validation* (LOOCV) is used to assess the accuracy of the generated meta-models.

This approach supposes to use a design of experiment $\mathcal{X}^{(N)} = \{(\boldsymbol{x}^{(1)}, f(\boldsymbol{x}^{(1)})), ..., (\boldsymbol{x}^{(N)}, f(\boldsymbol{x}^{(N)}))\}$ of N observations to estimate the coefficients $\beta_{\boldsymbol{\alpha}}$. The purpose of this communication is to present a novel approach to select iteratively the points of the design of experiments in order to estimate as fast as possible the 95%-quantile of the output distribution of Y. Let us consider a meta-model $\hat{M}$ constructed from the design $\mathcal{X}^{(N)}$ and constituted of the polynomials of a given set $\mathcal{A} \subset \mathbb{N}^m$:

$$\hat{M}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \beta_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \tag{2}$$

The principle of the proposed approach is first to frame and assess the variability of the meta-model with bootstrap resampling [2]. The Bootstrap uses a resampling of the design of experiments

and provides a framing of the values of the meta-model for all $\boldsymbol{x} \in \mathbb{X}$. Let us consider a set of B bootstrap resampling $\{\mathcal{X}_k^{*(N)},\ k \in [\![1, B]\!]\}$ of the design of experiments $\mathcal{X}^{(N)}$ and their corresponding meta-models $\{\hat{M}_k^*\}$. For $\boldsymbol{x} \in \mathbb{X}$, a α-confidence interval of $\hat{M}(\boldsymbol{x})$ can be extracted from the B corresponding Bootstrap meta-models:

$$\hat{M}^*_{[\alpha/2]}(\boldsymbol{x}) \leq M(\boldsymbol{x}) \leq \hat{M}^*_{[1-\alpha/2]}(\boldsymbol{x}) \tag{3}$$

Here the set $\{\hat{M}_k^*(\boldsymbol{x})\}$ is sorted in increasing order for each $\boldsymbol{x} \in \mathbb{X}$. Then we can define two confidence function L_α and U_α associating to each $\boldsymbol{x} \in \mathbb{X}$ the bounds of the inequality of Eq. (3), respectively. The following inequality can therefore be deduced:

$$L_\alpha(\boldsymbol{x}) \leq M(\boldsymbol{x}) \leq U_\alpha(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \mathbb{X} \tag{4}$$

From the latter equation, we can define a confidence interval of the 95%-quantile of the output variable:

$$q_{95_{L_\alpha}} \leq q_{95} \leq q_{95_{U_\alpha}} \tag{5}$$

Here $q_{95_{L_\alpha}}$, q_{95} and $q_{95_{U_\alpha}}$ are respectively the 95%-quantiles of $L_\alpha(\boldsymbol{X})$, $M(\boldsymbol{X})$ and $U_\alpha(\boldsymbol{X})$. In this communication, considering $\alpha = 0.05$, the area of selection for the new experiment of the design $\mathcal{X}^{(N+1)}$ is:

$$\mathcal{A}^{(N+1)} = \{\boldsymbol{x} \in \mathbb{X},\ U_\alpha(\boldsymbol{x}) \geq q_{95_{L_\alpha}} \text{ and } L_\alpha(\boldsymbol{x}) \leq q_{95_{U_\alpha}}\} \tag{6}$$

Then additional the point is selected in $\mathcal{A}^{(N+1)}$ following a maxmin criterion. The procedure is iteratively repeated in order to reduce the range of the inequality of Eq. (5) until this range is lower than a given stopping value.

This approach is illustrated with several benchmark functions used in the literature and with a dosimetry example aiming at assessing the influence of morphological parameters on the human exposure to electromagnetic fields.

References

[1] Géraud Blatman and Bruno Sudret. Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics*, 230(6):2345–2367, 2011.

[2] Bradley Efron and Gail Gong. A Leisurely Look at the Bootstrap , the Jackknife , and Cross-Validation. *The American Statistician*, 37(1):36–48, 1983.

Short biography – Engineer of Ecole Centrale de Nantes, after a master in Signal Processing, I am currently involved in the PhD thesis: "Statistical Analysis of People Exposure via the Numerical Dosimetry and the Design of Experiments" in Orange Labs and Université Paris-Est. The main purpose of this thesis is to statistically characterize the exposure induced by wireless communication systems.

Polynomial Chaos applied to the exposure assessment of child to Radio-Frequency field emitted by tablet devices

I. Liorni
Politecnico di Milano

Supervisor(s): Prof. Paolo Ravazzani (Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni IEIIT-CNR) and Dott.ssa Marta Parazzini (Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni IEIIT-CNR)

Ph.D. expected duration: 2012-2015

Adress: Piazza Leonardo da Vinci,32, 20133 Milano (Italy)

Email: ilaria.liorni@polimi.it

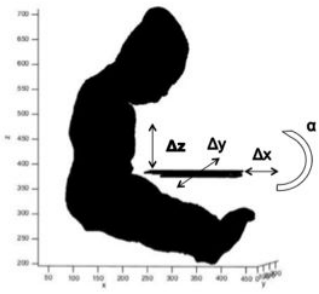
Abstract:

Introduction The public concern toward the exposure to Radio-Frequency electromagnetic fields (RF-EMF) exists despite the intensive use of wireless telecommunication systems. The assessment of real exposure scenarios is still an open issue, due to the variability of the input parameters that influence the exposure itself (e.g. the source design, the orientation of the incident field, the morphology of the subject exposed). Classical electromagnetic computational techniques are typically leading to highly time consuming simulations, if the variation of these parameters is taken into account using method such as Monte Carlo. Polynomial Chaos (PC) theory is a promising method to assess the variability of exposure at a lower computational cost [3]. In this study, PC theory has been applied to analyze the exposure of a 1-year-old baby to a 3G tablet emitting at 1940 MHz. The exposure has been characterized in terms of Specific Absorption Rate (SAR). A PC expansion has been built to estimate the whole-body SAR and the SAR in the brain, separately, to assess the variability of the child exposure with the change in the tablet position.

Material and methods The Polynomial Chaos is a spectral method and consists in the approximation of the system output in a suitable finite - dimensional basis $(\psi_j(\mathbf{X}))_{0\leq j\leq P-1}$ made of orthogonal polynomials. A truncation of this polynomial expansion can be performed:

$$Y = M(\mathbf{X}) = \sum_{k=0}^{P-1} a_j \psi_j(\mathbf{X}) \quad (1)$$

where Y is the system output (in this case the whole-body SAR and the SAR in the brain), $\mathbf{X}$ is the random input vector, and the a_j are the coefficients to be estimated. In this case, the input vector $\mathbf{X}$ is made of 4 independent parameters, supposed to be uniformly distributed and representing the translation of the tablet along the axis x,y and z and the rotation of the tablet in the xz-plane (Fig.1). The polynomial basis has been built using Legendre polynomials [4] and the a_j coefficients have been estimated by regression and Least Angle Regression (LAR) algorithm [1]. The best solution among the ones generated by LAR has been chosen through a leave-one-out (LOO) cross-validation. The observations $\mathbf{Y_0}$, used in LAR algorithm, have been estimated through computational dosimetry (Finite Difference Time Domain Method (FDTD)) by means of a numerical model of the child [2]. Each simulation lasts about 6 hours of computational time. The corresponding input $\mathbf{X_0}$ was generated by a Quasi Monte-Carlo method based on the Sobol function. The PC expansions, achieved by this procedure, have been validated by calculating the percentage mean square error (pMSE) on a validation set $\mathbf{Y_{val}}$ different from $\mathbf{Y_0}$. The PC expansions of the whole-body SAR and the SAR in the brain of the baby have been built by 60 observations with a pMSE equal to 0.02% and 0.27%, respectively, calculated on a set $\mathbf{Y_{val}}$ made of 30 observations. The moments of the first order have been analytically estimated through the



Input parameters	Variation
X1= Δx	[-10,20] mm
X2= Δy	[-40,40] mm
X3= Δz	[0,40] mm
X4= α	[0°,20°]

Figure 1: Baby sitting in the reference position with the 3G tablet at 20 cm from the eyes. In the table the input parameters and the corresponding variations are indicated

PC coefficients and the variability of the exposure has been assessed in terms of coefficient of variation (CV defined as the ratio of the standard deviation to the mean value). Finally, a global sensitivity analysis, expressed in terms of single-effect and total Sobol indices calculated from PC expansions, has been carried out to evaluate which input parameters influence more the exposure.

Results The CV was found up to 6.4% and 20% in the case of the whole-body SAR and the SAR in the brain, respectively. The global sensitivity analysis shows that parameter X1 (translation along x-axis) influences the variance of the whole-body SAR with a single-effect index of 46%. Furthermore total Sobol indices significantly higher than the single-effect indices are observed for X3 (translation along z-axis) and X4 (rotation α) meaning an interaction with the other parameters. Finally, the parameter X4 strongly influences the SAR in the brain with a single-effect index equal to 70%.

Conclusions In this work Polynomial Chaos (PC) has been used to assess the variability of the exposure of a baby to a 3G tablet at RF. PC resulted an efficient method to build accurate approximations of the SAR in the whole-body and in the brain with only 60 observations and with an error significantly lower than 0.5%. Variations of the RF exposure due to the change in the tablet position are up to 20% in the brain, in which the variation is mostly due to the rotation of the tablet in the xz-plane.

References

[1] B. Efron and et al. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[2] C. Li and et al. Generation of infant anatomical models for evaluating electromagnetic field exposures. *Bioelectromagnetics*, 36:10–26, 2015.

[3] J. Wiart and et al. Handle variability in numerical exposure assessment: the challenge of the stochastic dosimetry. *7th European Conference on Antennas and Propagation (EUCAP)*, 2013.

[4] D. Xiu and et al. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.

Short biography – Ilaria Liorni graduated in Biomedical Engineering at Sapienza Universita' di Roma. Since 2012 she is PhD student in Bioengineering at the Politecnico di Milano and she is associated to the Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni IEIIT (Consiglio Nazionale delle Ricerche CNR). Her PhD interests are about the investigation of deterministic and stochastic methods to assess the human exposure to electromagnetic fields (EMF).

Correspondence Between Gaussian Processes with Inequality Constraints and Constrained Splines

H. Maatouk
École Nationale Supérieure des Mines, Saint-Étienne, France

Supervisor(s): Xavier Bay (ENSM-SE), Laurence Grammont (ICJ, Lyon 1), Yann Richet (IRSN, Paris) and Olivier Roustant (ENSM-SE)

Ph.D. expected duration: 2012-2015

Adress: 158 cours Fauriel, 42 023 St-Etienne cedex 2, France

Email: hassan.maatouk@mines-stetienne.fr

Abstract: In statistical emulation of computer experiments, the physical system (computer model output) may be known to satisfy inequality constraints with respect to some or all input variables. A new methodology to incorporate both equality and inequality constraints into a Gaussian process emulator has been developed in [3]. The main idea of this approach is based on an uniform pathwise approximation of the original Gaussian process Y by a finite-dimensional Gaussian process Y^N of the form:

$$Y^N(\boldsymbol{x}) = \sum_{j=0}^{N} \xi_j \phi_j(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^d \tag{1}$$

where $\xi = (\xi_0, \ldots, \xi_N)^\top$ is a centered Gaussian vector with carefully chosen covariance matrix Γ_N and deterministic basis functions $(\phi_j)_j$. We show in [3] that $(\phi_j)_j$ can be specified such that the inequality constraints of Y^N are *equivalent* to constraints on the coefficients. The important point of such approach is that the inequality constraints are reduced to a finite number of constraints and satisfied in the whole domain. As a result, the problem is reduced to simulate a truncated Gaussian vector restricted to a convex set with known algorithms (see e.g. [2]). In order to investigate the performance of the proposed model, some conditional simulations with inequality constraints such as boundary, monotonicity or convexity conditions in one and two dimensions are given (see Figure 1). In real application, the parameters estimation should be investigated and Cross Validation techniques are used.

In the second part of the talk, the convergence of the proposed model is investigated and the following optimization problem is considered:

$$\inf_{h \in H \cap I \cap C} \|h\|_H^2, \tag{P}$$

where H is an Hilbert space with norm $\|.\|_H$, I is the space of interpolating functions and C is a closed convex set of $\mathbb{R}^d$ corresponding to inequality constraints. We prove that the mode of the conditional Gaussian process Y^N (maximum a posteriori) converges uniformly to the solution of problem (P) called the constrained smoothing spline (see below Figure 2). This result can be seen as an extension of the correspondence established by Kimeldorf and Wahba [1] between Bayesian estimation on stochastic process and smoothing by splines.

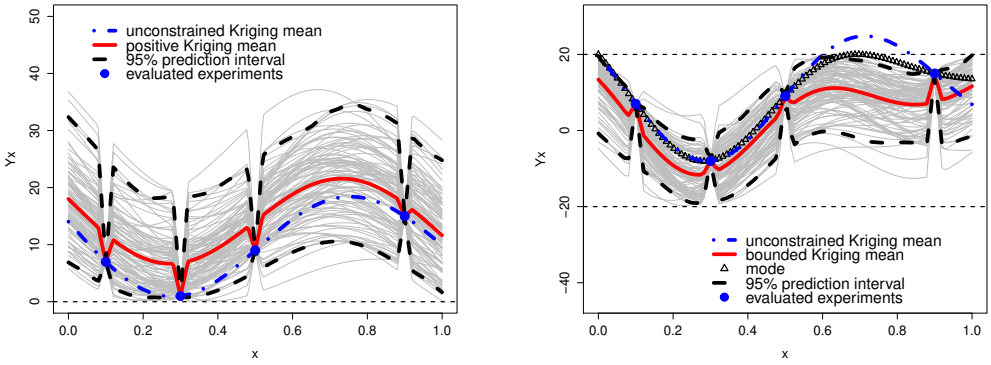
References

[1] George S. Kimeldorf and Grace Wahba. A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 04 1970.

[2] Hassan Maatouk and Xavier Bay. A New Rejection Sampling Method for Truncated Multivariate Gaussian Random Variables Restricted to Convex Sets. hal-01063978, September 2014.

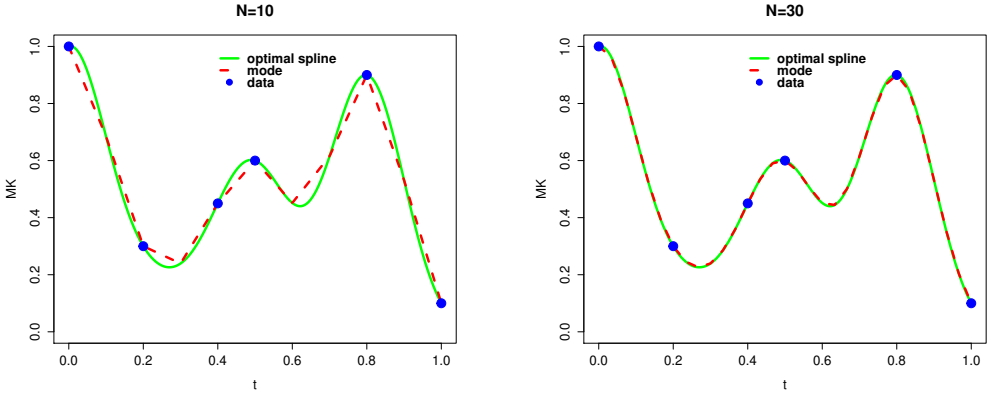
[3] Hassan Maatouk and Xavier Bay. Gaussian Process Emulators for Computer Experiments with Inequality Constraints. hal-01096751, December 2014.

[4] Edward J. Wegman and Ian W. Wright. Splines in Statistics. *Journal of the American Statistical Association*, 78(382):pp. 351–365, 1983.



(a) (b)

Figure 1: Simulated paths (gray lines) drawn from the conditional GP respecting positivity constraints Figure 1a and boundary constraints Figure 1b.



(a) (b)

Figure 2: The green line represents the interpolation spline with inequality constraints (boundary constraints). The red dashed line is the mode (maximum a posteriori) of the finite-dimensional GP conditionally to both equality conditions and inequality constraints. It respects boundary constraints and coincides quickly with the green line when N is large enough.

Short biography – Hassan Maatouk is a third year PhD student in applied mathematics at the ENSM-SE (French engineering school). My work is a collaboration between ENSM-SE, Lyon 1 University and IRSN (Paris). It focuses on incorporating both data interpolation and inequality constraints into a Gaussian process emulator.

Polynomial chaos expansions for time-dependent problems

C.V. Mai
ETH Zürich, Zürich, Switzerland

Supervisor(s): B. Sudret (ETH Zürich)

Ph.D. expected duration: 2012-2016

Adress: ETH Zürich, Institute of Structural Engineering, Chair of Risk, Safety & Uncertainty Quantification, Stefano-Franscini-Platz 5, CH-8093 Zürich, Switzerland

Email: mai@ibk.baug.ethz.ch

Abstract:

Due to the increasing complexity of physical systems and associated computational models, uncertainty quantification has become a key topic in modern engineering. One of the objectives of its formulation is to propagate uncertainties from the stochastic input parameters of the model to the output quantities of interest. For this purpose, polynomial chaos expansions (PCE) have been widely used as surrogate models (or metamodels) that substitute computationally expensive ones.

PCE, however, face challenges when applied to time-dependent problems, *e.g.* involving structural, fluid dynamics or chemical systems. In general, the PCE of a time-dependent quantity reads:

$$x(t, \boldsymbol{\xi}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} x_{\boldsymbol{\alpha}}(t) \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) \tag{1}$$

in which $\boldsymbol{\xi}$ represents the random input parameters, $\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ are the polynomial chaos basis and $x_{\boldsymbol{\alpha}}(t)$ are the time-dependent deterministic coefficients, which are computed point-wise in time. The greatest challenge hindering the use of PCE is the decrease of its accuracy over time [1, 2], which is illustrated in Figure 1.

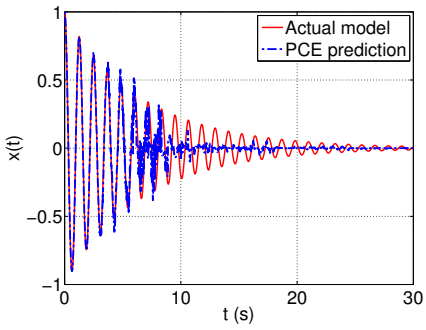


Figure 1: Prediction of the displacement $x(t)$ of a non-linear Duffing oscillator by PCE *vs.* the actual response trajectory.

To overcome such limitation, [2] proposed an *intrusive* time-transform approach in which the responses are represented in a new time scale. This approach is, however, not applicable in case the governing equations of the considered system are not known explicitly. In this work, we introduce a *non-intrusive* approach that allows the effective use of PCE for time-dependent models which are considered as "black-boxes". It relies on a *non-intrusive stochastic time-transform* of

the response trajectories which aims at maximizing the similarities in frequency and phase content of the sampled time-histories. The linear stochastic time-transform operator reads:

$$\tau = kt + \phi \tag{2}$$

where k and ϕ are the parameters governing the scaling and shifting of the original time t. k and ϕ are determined by maximizing the similarity between the considered response trajectory and a selected reference counterpart. We propose a measure to quantify the similarity between distinct time-histories in terms of frequency and phase content, which reads:

$$g(k,\phi) = \frac{\left|\int_0^T x_i(k\,t+\phi)x_r(t)\mathrm{d}t\right|}{\|x_i(k\,t+\phi)\|\|x_r(t)\|}, \tag{3}$$

where $\int_0^T x_1(t)x_2(t)\mathrm{d}t$ is the inner product of two time-histories $x_1(t)$ and $x_2(t)$; $\|\cdot\|$ is the associated L^2-norm, *i.e.* $\|x_1(t)\| = \sqrt{\int_0^T x_1^2(t)\mathrm{d}t}$. Polynomial chaos expansions are used to model the responses of the considered system in the transformed time τ. The responses are subsequently mapped back onto the original time t. Figure 2 depicts the time-history response of a nonlinear damped Duffing oscillator predicted by time-transform PCE versus the actual response. Other examples in mechanical and chemical engineering are also used to demonstrate the effectiveness of the proposed approach.

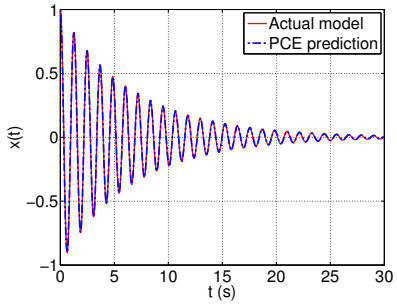


Figure 2: Prediction of the displacement $x(t)$ of a non-linear Duffing oscillator by time-transform PCE *vs.* the actual response trajectory.

References

[1] D Ghosh and G Iaccarino. Applicability of the spectral stochastic finite element method in time-dependent uncertain problems. *Annual Research Briefs of Center for Turbulence Research*, pages 133–141, 2007.

[2] O Le Maître, L Mathelin, O Knio, and M Hussaini. Asynchronous time integration for polynomial chaos expansion of uncertain periodic dynamics. *Discret. Contin. Dyn. Sys. - Series A (DCDS-A)*, 28(1):199–226, 2010.

Short biography – Chu V. Mai received his bachelor in civil engineering at National University of Civil Engineering in Hanoi, Vietnam in 2011. He obtained a Master degree in Materials and Structures from the Paris-Est University in 2012. His PhD topic focuses on the use of sparse polynomial chaos expansions in structural dynamics with applications to earthquake engineering. His research interests also include reliability and global sensitivity analyses.

Learning non-stationary zones with warped Gaussian processes

S. Marmin
Univeristy of Bern

Supervisor(s): David Ginsbourger (University of Bern), Jacques Liandrat (École Centrale de Marseille) and Dr. Jean Baccou (Institut de Radioprotection et de Sûret Nucláire)

Ph.D. expected duration: 2014-2017

Adress: Universität Bern, IMSV, Alpeneggstrasse 22, CH-3012 Bern, Switzerland

Email: sebastien.marmin@irsn.fr

Abstract:

When a computer code is expensive to evaluate, it seems essential for the user to look for a design of experiment adapted to its needs, anticipating the response of the code. To predict the response, one can build a model extrapolating the initial data in unexplored areas. Here we consider models substituing the unknown function by an interpolating Gaussian process (kriging). The covariance function of the Gaussian process is often assumed stationary. This hypothesis can lead to poor predictions if the response behaves heterogeneously across the input space. This kind of behaviour happens in many nuclear-safety studies conducted by IRSN which motivate this work. The first part of the presentation is dedicated to the comparison and validation of non-stationary Gaussian process models. The data comes from a IRSN computer code simulating the fracture dynamics of heterogeneous materials and used to study the mechanical ageing of nuclear power plants. Several known non-stationary modeling aproaches are reviewed : source-space warping, Gaussian processes generated by convolution and space-dependent combinations of different Gaussian processes. The second part is about adaptive design of experiment for the exploration of high variation regions. We define and compare criteria which detect non-stationary zones. In conclusion, we present progresses made on the nuclear-safety test case.

Short biography – Before my PhD, I have studied applied mathematics at École des Mines de Saint-Étienne. The first year of my PhD takes place at Bern University. Then I will continue my works in a laboratory of the IRSN, my employer, at Cadarache (PACA, France) for two years.

An analytic comparison of regularization methods for Gaussian Processes

H. Mohammadi, R. Le Riche, E. Touboul, X. Bay, N. Durrande
École Nationale Supérieure des Mines de Saint-Étienne (EMSE)

Supervisor(s): Prof. Le Riche (CNRS and EMSE), Prof. Touboul (EMSE)

Ph.D. expected duration: 2013-2016

Adress: 158 Cours Fauriel, 42023, Saint-Étienne, France

Email: hossein.mohammadi@emse.fr

Abstract:

Conditional Gaussian Processes (GPs), also known as kriging models, are often used to predict the output of a parameterized deterministic (non noisy) experiment. They have many applications in the field of Computer Experiments, in particular to perform sensitivity analysis [3], adaptive design of experiments [1], and global optimization [2]. Nearly all of the applications of GPs to Computer Experiments require the inversion of a covariance matrix. In practice, this matrix is often ill-conditioned and therefore not numerically invertible. It happens when observed points are repeated, or even are close to each other, or when the covariance function makes the information provided by observations redundant.

In the literature, various strategies have been proposed to avoid degeneracy of the covariance matrix. The two most classical regularization methods are *i*) adding a small positive constant to the diagonal (which we will refer to with a slight abuse of language as "nugget" regularization) [4] and *ii*) pseudoinverse (PI) [2] in which singular values smaller than a threshold are zeroed. These two techniques have a wide range of applications since they can be used a posteriori in computer experiments algorithms without major redesign of the methods. Today, there is still a need to better understand and compare the effects of these techniques on GPs.

This work considers GP models in the context of deterministic experiments where interpolation is sought but classical models would need regularization techniques.

In a first part, we provide algebraic calculations which allow comparing PI and nugget regularizations. The focus is on interpolation properties of the conditional GPs when the observed points tend towards each other, i.e., become *redundant*. The analysis is made possible by approximating ill-conditioned covariance matrices with the neighboring truly non invertible covariance matrices. If the redundant points have different outputs, we prove the following results :

- PI regularization averages the output values and has null variance at the redundant points.
- Nugget regularization tends to the same behavior when the nugget magnitude decreases.
- Maximum likelihood estimation of the nugget term may yield large amplitudes, which degrade the interpolation quality. In this sense, in a noise free context, pseudoinverse or fixed small nugget values should be preferred to maximum likelihood estimation of the nugget.

We suggest how to set nugget and PI threshold values based on the condition number of the covariance matrix (the ratio of the largest to the smallest eigenvalue).

In a second part, we propose a GP model with improved interpolation properties at redundant points. It is a *distribution-wise* GP model : the trajectories pass through data points that are uniquely defined (non-redundant), therefore having a null variance there; at redundant points

with different outputs, the GP mean and variance are the empirical average and variance of the outputs, respectively. This model is obtained by conditioning on probability distributions instead of data point(s). The distribution-wise GP can be seen as new a posteriori regularization technique through clustering of the points.

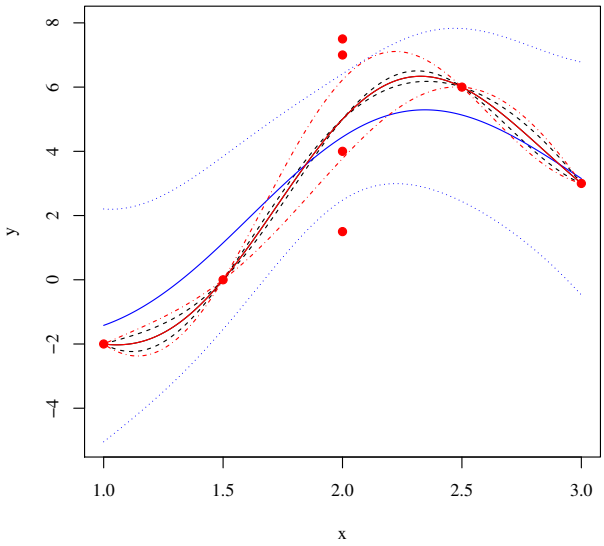


Figure 1: Comparison of kriging means (the solid lines) and kriging means (+/-) std dev.; maximum likelihood nugget regularization (dotted lines), PI regularization (dashed lines) and distribution-wise GP (DGP, dash-dotted lines). The bullets are data points. PI and DGP pass through the mean of outputs at redundant points while nugget does not. Contrarily to PI, DGP captures the empirical variance.

References

[1] Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vázquez. Sequential Design of Computer Experiments for the Estimation of a Probability of Failure. *Statistics and Computing*, 22(3):773–793, 2012.

[2] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[3] Jeremy E. Oakley and Anthony O'Hagan. Probabilistic Sensitivity Analysis of Complex Models: A Bayesian Approach. *Journal of the Royal Statistical Society, Series B*, 66(3):751–769, 2002.

[4] Olivier Roustant, David Ginsbourger, and Yves Deville. DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization. *Journal of Statistical Software*, 51(1):1–55, 2012.

Short biography – Hossein Mohammadi received the B.S. and M.S. degrees in industrial engineering from Amirkabir University of Technology, Tehran, Iran, in 2008 and 2011. He is currently a PhD candidate at Ecole des Mines de Saint-Etienne, France. His research interests are statistical learning by Gaussian processes, global optimization of expensive functions, and their applications in engineering and other sciences.

Excess probability and quantile estimation for monotone codes

V. Moutoussamy
University Paul Sabatier - Toulouse

Supervisor(s): F. Gamboa (University Paul Sabatier Toulouse), T. Klein (University Paul Sabatier Toulouse), N. Bousquet (EDF) and B. Iooss (EDF)

Ph.D. expected duration: 2012-2015

Adress: 6, quai Watier - 78401 Chatou

Email: vincent.moutoussamy@edf.fr

Abstract: This presentation focus on the theoretical convergence properties of the estimators produced by a numerical exploration of a monotonic function with multivariate random inputs $\mathbf{X}$. The quantity to be estimated is a probability typically associated to an undesirable (unsafe) event and the function is usually implemented as a computer model g. That probability is defined by $p = \mathbb{P}(g(\mathbf{X}) \leq z)$ where z is a threshold given by the user. The estimators produced by a Monte Carlo numerical design are two subsets of inputs leading to safe and unsafe situations, the measures of which can be traduced as deterministic bounds for the probability.

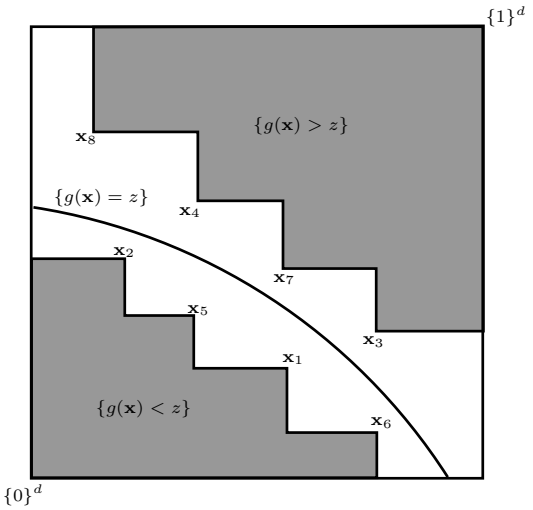


Figure 1: From a set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_8\}$, the two dominated space represented in gray can be deduced. In these two subsets, the sign of g is perfectly known.

A major consequence, a consistent estimator of the (limit state) surface separating the subsets under isotonicity and regularity arguments can be built, and its convergence speed can be exhibite: almost surely

$$d_H(F_n, \mathbb{U}^-) = O\left(\left(\frac{log(n)}{n}\right)^{1/d}\right),$$

with d_H the Hausdorff distance and F_n an estimator of the unsafe space. Such a situation can be encountered too when computing a Pareto frontier in multi-objective optimization. This estimator is build by aggregating semi-supervized binary classifiers chosen as constrained Support Vector

Machines. Numerical experiments conducted on toy examples highlight that they work faster than recently developed monotonic neural networks with comparable predictable power, and are, therefore, more adapted to situations when the computational time is a key issue.

In a second part, our presentation focus on quantile estimation under constraint monotonicity defined by $q = \inf\{z \in \mathbb{R},\ \mathbb{P}(g(\mathbf{X}) \leq z) \geq p\}$, where p is given. Under a geometrical constraint not too restrictive, a quantile can be bounded from a given set of points. Like for probability estimation, the input space can be constrained. Adapting a probability estimator in a sequential framework and using the bouding method, an adaptive quantile estimator can be constructed. Finally, this estimator is tested on numerical example.

References

[1] N. Bousquet. Accelerated Monte Carlo estimation of exceedance probabilities under monotonicity constraints. *Annales de la Faculté des Sciences de Toulouse*, 21:557–591, 2012.

[2] H. Daniels and M. Velikova. Monotone and partially monotone neural networks. *Neural Networks, IEEE Transactions on*, 21(6):906–917, 2010.

Short biography – After a Master in applied mathematics, Vincent Moutoussamy begun a thesis, a partnership between Electricité De France (EDF) and the University Paul Sabatier Toulouse, on monotonic structural reliability methods.

Sensitivity analysis of computer codes with functional inputs

S. Nanty
CEA, DEN, DER, F-13108, Saint-Paul-lez-Durance, France
Université Joseph Fourier, Grenoble

Supervisor(s): C. Prieur (Laboratoire Jean Kuntzman), C. Helbert (Ecole Centrale Lyon), A. Marrel (CEA) and N. Pérot (CEA)

Ph.D. expected duration: 2012-2015

Address: CEA Cadarache, F-13108, Saint-Paul-lez-Durance, France

Email: simon.nanty@cea.fr

Abstract:

In the framework of industrial risk assessment studies, the reliability of a component is evaluated during accidental conditions. A numerical model (called C_1 code) takes as inputs thermal-hydraulic (T-H) parameters and provides dependent temporal variables which describe the T-H loading of the component. Then, a numerical model for the mechanical analysis of components and structures (called hereafter C_2 code) takes as input the T-H loading and also thermal-mechanical (T-M) parameters. C_2 calculates the breaking strength of the component and the thermo-mechanical actual applied load. From these two elements, a safety criterion is deduced. This complete workflow is drawn in Figure 1.

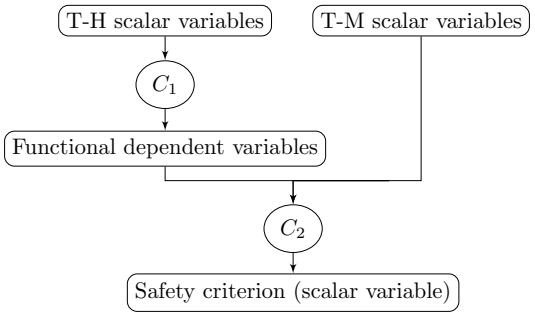


Figure 1: Workflow of the two chained computer codes.

The component behavior predicted by the workflow depends on many uncertain parameters, related to the initial plant conditions or to the safety system characteristics. C_1 code takes as inputs scalar T-H uncertain parameters and C_2 code takes as inputs C_1 code results and independent scalar T-M uncertain parameters. All these T-H and T-M parameters can greatly affect the code forecasts. Our objective is to quantify the influence of these uncertain parameters through sensitivity analysis (SA), and more specifically the influence of the group of T-H parameters compared to the influences of T-M parameters. Among all existing SA methods, Sobol indices [5] are discussed here. This widely used variance-based method quantifies and ranks the contributions of the input variables and their interactions to the code output uncertainty. Sobol indices are defined for independent scalar input parameters but their definition remains valid for functional input variables and for grouped variables (Jacques et al. [3]). However, such quantitative SA methods need thousands of computer code evaluations, and C_1 code is time expensive (only a few hundred evaluations can be done), while C_2 is faster. Consequently, performing SA directly on the two chained codes is intractable.

To cope with this limitation, we propose a methodology which combines an uncertainty characterization of the outputs of the slowest code (C_1) and a direct intensive SA of the faster code (C_2). Indeed, as the influence of C_1 code outputs traduces the influence of T-H inputs, the SA can be performed on C_2 considering as inputs the T-M parameters and C_1 outputs. This SA requires to know the probability density function of C_1 functional outputs which has to be estimated based on available realizations. This functional input characterization is done in two steps. First, the functional variables are decomposed simultaneously on a functional basis. The decomposition basis is truncated and a small number of basis functions is selected. The variables are summarized by their coefficients on these basis functions. The decomposition method must preserve both the dependence between the function variables and the link between functional variables and C_2 code output. To this mean, a simultaneous version of Partial Least Squares decomposition (Wold [6]) has been developed. The probability density distribution of the decomposition coefficients is then modeled by a Gaussian mixture model. In order to estimate the Gaussian mixture model parameters, we propose a sparse version of the Expectation-Maximization algorithm (Dempster et al. [1]), based on a Lasso penalization. Through these two steps, a statistical quantification of the uncertainties on the dependent functional random variables is achieved. The whole quantification methodology is fully described in [4]. With this methodology, the uncertainties of the functional inputs of the C_2 computer code are quantified. It is then possible to perform SA on the C_2 code. For this, we propose to estimate Sobol indices with the method for grouped variables based on replicated Latin hypercube designs and developed by Gilquin et al. [2].

References

[1] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[2] Laurent Gilquin, Clémentine Prieur, and Elise Arnaud. Replication procedure for grouped Sobol' indices estimation in dependent uncertainty spaces. 2014.

[3] Julien Jacques, Christian Lavergne, and Nicolas Devictor. Sensitivity analysis in presence of model uncertainty and correlated inputs. *Reliability Engineering & System Safety*, 91(10-11):1126 – 1134, 2006.

[4] Simon Nanty, Céline Helbert, Amandine Marrel, Nadia Pérot, and Clémentine Prieur. Uncertainy quantification for functional dependent random variables. Submitted to Statistics and Computing, 2014.

[5] Ilya M. Sobol. Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.

[6] Herman Wold. *Estimation of Principal Components and Related Models by Iterative Least squares*, pages 391–420. Academic Press, 1966.

Short biography – After a master's degree in applied mathematics in Saint-Etienne, I began my PhD in 2012 in collaboration with Grenoble University and the Commissariat à l'Energie Atomique. This PhD thesis is funded by the CEA, and its objective is to quantify the uncertainties associated to a computer code with functional inputs.

Kernels on the unit disk for spatial uncertainty assessment

Espéran Padonou
Mines Saint-Étienne - STMicroelectronics

Supervisor(s): Olivier Roustant (Mines Saint-Étienne), Jakey Blue (Mines Saint-Étienne) and Hugues Duverneuil (STMicroelectronics)

Ph.D. expected duration: 2013-2016

Adress: padonou@emse.fr

Email: padonou@emse.fr

Abstract: In semiconductor industry, Integrated Circuits are manufactured on the surface of a circular slice of semiconductor material called wafer. For quality inspection, several measurements such as thickness, depth and width, are taken on each wafer. These data are collected on a set of predefined points. Because these locations are not uniformly affected by process conditions, the associated measurements will not be identically distributed. As consequence, the risk of default can notably vary from an area to another on the same product.

Among the models that exist in the litterature of response surface methodology, few were focused on the case of circular domains. In particular, Zernike regression [5] is a parametric model adapted for circular shapes. As for non-parametric methods, spline smoothing [1] and more recently Kriging [3] led to good results in the framework of quality control in microelectronics. Kriging is an interpolation method which infers the expected value of a Gaussian Random Field conditionally to observations. It estimates the response value at any location and provides the corresponding uncertainty (Kriging standard deviation). Its accuracy strongly depends on the choice of the covariance function or kernel. Usually, kernels based on Euclidian distances such as the Gaussian or Matérn functions are used. In Cartesian coordinates, they are written:

$$k\left((x,y),(x',y')\right) = \Phi\left(k_1\left(x,x'\right),k_2\left(x,x'\right)\right)$$

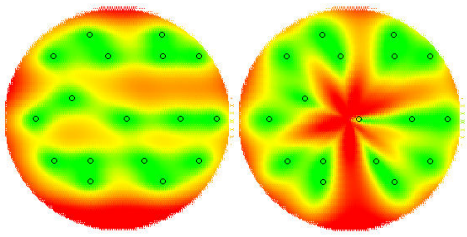
Where Φ is an operator, such as: product, addition or ANOVA [4]. A matérn$_{\frac{5}{2}}$ is chosen for k_1 and k_2. In this study, we first define new kernels in polar coordinates:

$$k\left((\rho,\theta),(\rho',\theta')\right) = \Phi\left(k_r\left(\rho,\rho'\right),k_a\left(\theta,\theta'\right)\right)$$

For k_r, the matérn$_{\frac{5}{2}}$ can be used too. The choice of the angular kernel k_a is justified via theoretical results on positive definite functions on the circle [2]. We show via simulations and industrial cases that the approach improves the kriging interpolation when the response includes angular variations (due to rotations for instance). Regression models and kriging with different kernels are compared (Table 1).

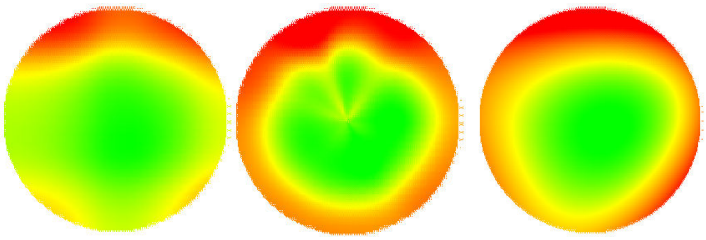
	Kriging with a Euclidian distance			Kriging with a distance using arcs			Regression
Type	Product *(a)*	Additive	ANOVA	Product	Additive	ANOVA *(b)*	Polynomial *(c)*
RMSE	51.9	55.5	54.4	59.9	47.0	46.5	62.9

Table 1: Results for different models based on the design presented in Figure 1. 64 new points serve as test set to compute the Root Mean Square Errors (RMSE). The polynomial regression is the least efficient. Kriging with the proposed kernel (b) leads to the best result.



Kernel a Kernel b

Figure 1: Kriging standard deviations for kernel a (using an Euclidian distance) and kernel b (using a distance based on the arc). Black circles represent the 17 design points. Kernel (a) leads to elliptical contours around the points of the design whereas the contours produced by kernel (b) look like pie chart sectors. In addition for this latter kernel, the frequency changes from the center to the boudary of the domain.



Kriging (kernel a) Kriging (kernel b) 3 order polynomial (c)

Figure 2: Predicted values for different models: adding a polynomial trend (c) does not improve the Kriging model. The new distance (kernel b) improves the estimation near the boundary.

References

[1] M. M. Gardner, J. C. Lu, R. S. Gyurcsik, J. J. Wortman, B. E. Hornung, H. H. Heinisch, E. A. Rying, S. Rao, J. C. Davis, and P. K. Mozumder. Equipment fault detection using spatial signatures. *IEEE Transaction on Components, Packaging, and Manufacturing Technology - Part C*, 20(3):295–303, 1997.

[2] Tilmann Gneiting. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4):1327–1349, 09 2013.

[3] Giovanni Pistone and Grazia Vicario. Kriging prediction from a circular grid: application to wafer diffusion. *Applied Stochastic Models in Business and Industry*, 29(4):350–361, 2013.

[4] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[5] Frederik Zernike. Diffraction theory of the cut procedure and its improved form, the phase contrast method. *Physica*, 1:689–704, 1934.

Short biography – Esperan Padonou is a PhD student in applied mathematics at Mines Saint-Etienne. His research is ongoing through a partnership with the company STMicroelectronics and consists in developing statistical solutions for risk assessment and control planning in semiconductor industry.

Propagation of imprecise probabilities using sparse polynomial chaos expansions

R. Schöbi
Chair of Risk, Saftey and Uncertainty Quantification, ETH Zurich, Switzerland

Supervisor(s): Prof. B. Sudret (ETH Zurich)

Ph.D. expected duration: 2013-2016

Adress: Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland

Email: schoebi@ibk.baug.ethz.ch

Abstract: In modern engineering, it is typical to model physical systems and processes with advanced computational models, such as finite element models, in order to assess their performance and optimize them. Moreover, awareness is growing on concepts like structural reliability and robust design, hence making the quantification and propagation of uncertainties a key issue.

Due to the high cost of evaluating repeatedly complex computational models, simulations with classical sampling techniques such as Monte Carlo (MC) are often intractable. In this context, meta-modelling techniques allow one to develop surrogate models from a limited number of runs of the original computational model. In non-intrusive approaches, the computational model is interpreted as a black-box function which maps the M-dimensional input space to a scalar output space. A popular approach hereof is Polynomial Chaos Expansions (PCE), which approximate the response of the computational model with a set of multivariate polynomials that are orthogonal with respect to the distribution of the input variables [3]. The coefficients of a PCE can be computed *e.g.* by generalized least-squares minimization [1].

The surrogate model is a fast-to-evaluate function which allows one to perform reliability analysis and design optimization at significantly reduced computational costs. In these analyses, the proper probabilistic description of the input variables is of utmost importance. Traditionally, the input variables are determined by statistical inference from experimental data assuming they follow certain probability distributions. The parameters of the probability distributions are typically determined by maximum likelihood estimation. The uncertainty on the parameters arising from the sparsity of the experimental data is often neglected. However, due to resource limitations in terms of costs, time and labour, it is common in real applications that only a limited amount of input measurements and/or imprecise data is available to estimate those distribution parameters.

Therefore, a general framework is needed to capture the uncertainty in the parameters of the probability distributions. Imprecise/mixed probability is a concept that describes input variables combining different types of uncertainty, such as epistemic (lack of data) and aleatory (natural variability). Among others, probability-boxes (p-boxes) [2] and the Dempster-Shafer's theory of evidence [5] have been proposed in order to characterize these uncertainties.

The computational burden of propagating imprecise probabilities is large because the various formulation leads to a more complex description of the input space. This is where meta-models come into play: their efficiency allows for fast propagation of uncertainty though the computational model and thus, a more efficient estimation of the quantity of interest in the output space.

In this work, we propose an algorithm which combines the framework of p-boxes with the method of sparse PCE in order to predict the p-box of the response quantity efficiently. Particularly, we focus on parametric p-boxes which is a special case of p-boxes. The parametric p-box assumes that an imprecise random variable X can be described by a probability distribution family whose parameters are known within certain ranges rather than having deterministic values. In this

model, the parameters represent the epistemic uncertainty, whereas the shape of the probability distribution represents the aleatory uncertainty. For a single parameter X, we define:

$$X: \; x \sim F_X = F_X(x|\boldsymbol{\theta}) \quad \text{where } \boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\Theta}}, \; \theta_i \in [\underline{\theta}_i, \overline{\theta}_i], \; i = 1, \ldots, n_{\boldsymbol{\theta}}, \tag{1}$$

where F_X denotes the cumulative distribution function of variable X and $[\underline{\theta}_i, \overline{\theta}_i]$ are the lower and upper bounds of its parameter θ_i. $\boldsymbol{\Theta}$ describes the collection of $\boldsymbol{\theta}_j$ for each X_j, $i = 1, \ldots, M_X$.

This separation of aleatory and epistemic uncertainty allows one to build a meta-model in an *augmented input space* containing the variable X and its uncertain probability distribution parameters $\boldsymbol{\theta}$, denoted by $\boldsymbol{Z} = \{X, \boldsymbol{\theta}\}$. At the same time, this increases the dimensionality of the input space of the meta-model. For instance if X is a Gaussian variable whose mean and standard deviation are defined by bounds then $\boldsymbol{Z} = \{X, \mu, \sigma\}$ has three components.

In the general case when $\boldsymbol{X}$ is a random vector with independent components, the dimensionality of the computational model is M_X, whereas the dimensionality of the augmented input space is $M_Z = M_X + |\boldsymbol{\Theta}| > M_X$. This means that several samples $\boldsymbol{Z}$ in the augmented input space will lead to the same corresponding sample points in the original input space X. This property can be exploited to generate artificial samples for the experimental design of the meta-model (so-called *phantom points*) that do not require additional runs of the original computational model. The induced phantom points increase the size of the experimental design and thus, improve the accuracy of the meta-model in the augmented input space [4].

Finally, the propagation of the input p-box is done with nested Monte-Carlo loops. The outer loop yields samples of the distribution parameters ($\boldsymbol{\theta}^* \in \mathcal{D}_{\boldsymbol{\Theta}}$) of the parametric p-box, whereas the inner loop yields samples of the probability distribution conditional on the sampled values of the parameters ($\boldsymbol{x} \sim F_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}^*)$). The boundaries of the p-box of the response variable Y are computed as the minimum/maximum values of the set of cumulative distribution families obtained by the sampled parameters $\boldsymbol{\theta}$ at each value of $y \in Y$.

The characteristics and capabilities of the proposed approach are analysed on benchmark analytical functions and realistic engineering systems. The results show that the combination of parametric p-boxes and PCE allows for an efficient prediction of the imprecise variable in the output of the computational model.

References

[1] G. Blatman and B. Sudret. Adaptive sparse polynomial chaos expansion based on Least Angle Regression. *J. Comput. Phys*, 230:2345–2367, 2011.

[2] Scott Ferson and Janos G. Hajagos. Arithmetic with uncertain numbers: rigorous and (often) best possible answers. *Reliab. Eng. Sys. Safety*, 85(1-3):135–152, July 2004.

[3] R. Ghanem and P. Spanos. *Stochastic Finite Elements : A Spectral Approach*. Dover Publications, Mineola, 2nd edition, 2003.

[4] R. Schöbi and B. Sudret. Propagation of uncertainties modelled by parametric p-boxes using sparse polynomial chaos expansions. In *Proc. 12th Int. Conf. on Applications of Stat. and Prob. in Civil Engineering (ICASP12), Vancouver, Canada*, 2015.

[5] G Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.

Short biography – R. Schöbi graduated in 2012 from ETH Zurich with a Master's degree in civil engineering where his Master's Thesis was entitled "Subset Simulation in Engineering Problems". Since May 2013, he is a Ph.D. student at the Chair or Risk, Safety and Uncertainty Quantification of ETH Zurich under the supervision of Prof. B. Sudret. His research topics include the quantification and propagation of epistemic uncertainty using mixed probabilistic/nonprobabilistic approaches and surrogate modelling using polynomial chaos expansions and Kriging.

Predicting outputs of a reservoir model with multi-fidelity meta-models

A. Thenon
IFP Energies nouvelles (IFPEN)

Supervisor(s): M. Le Ravalec (IFPEN) and V. Gervais (IFPEN)

Ph.D. expected duration: 2014-2017

Adress: IFP Energies Nouvelles - 1 & 4, avenue de Bois-Prau - 92852 Rueil-Malmaison Cedex

Email: arthur.thenon@ifpen.fr

Abstract: In the oil and gas industry, numerical 3D models describing the spatial distribution of reservoir properties are usually built to estimate recoverable reserves and help to manage the future development of the field. However, these models depend on parameters that are generally uncertain. To make them as representative as possible, data collected on the field are considered. Constraining reservoir models to production data is usually referred to as history matching. During this key process, a minimization is performed to reduce the mismatch between the measured data and the corresponding simulated values. The data mismatch is usually quantified from a function called objective function. The minimization process requires simulating fluid flow behavior for many reservoir models, which can be very time-consuming depending on the size of the model.

To handle this difficulty, we can build a meta-model that approximates the response of interest (objective function, simulated dynamic property) from a limited number of simulations. The meta-model can then be used to predict the response at new points of the parameter space, and thus limit the calls to the fluid-flow simulator. Several methods have been proposed to build such meta-models. In particular, kriging proved to be efficient in reservoir applications [2]. To further reduce computation times, we propose to investigate a more complex type of meta-models proposed in [4]. This model is able to integrate information collected at various levels of information: multi-fidelity meta-models. Even though less informative fluid-flow simulations considering coarser grids for the reservoir model are used, we expect to perform faster.

Multi-fidelity meta-modeling [4] is rooted in co-kriging and makes use of information related to the response of interest available at various levels of resolution to build a meta-model approximating the response on the finest level. This approach aims at building meta-models of good quality while limiting the calls to the more expensive simulators.

A synthetic reservoir model derived from PUNQ-S3 [3] is considered here to investigate the potential of multi-fidelity meta-models in reservoir applications. Two levels of resolution, defined by the spatial discretization of the reservoir domain used to perform fluid-flow simulations, are considered. The fine level is given by a grid of $57 \times 84 \times 5$ blocks and the coarse level by a grid of $19 \times 28 \times 5$ blocks. A fluid-flow simulation performed on the coarse grid is then 18 times faster on average than a simulation on the fine level. Seven uncertain parameters related to the reservoir model are considered: three residual saturations, three permeability multipliers (one per layer) and one permeability multiplier for the aquifer.

Multi-fidelity meta-models were built to approximate the dynamic responses at wells and the objective function (least-square formulation) using [5]. To approximate the time-dependent responses at wells, the reduced-basis approach based on PCA decomposition used in [1] was considered and coupled to the multi-fidelity approach. Kriging-based models were also computed for comparisons. The predictivity of the resulting meta-models was assessed through the computation of the Q2 coefficient on an independent set of 200 reservoir models.

When focusing on dynamic responses at wells, the meta-model quality was shown to depend on the considered output and times. Nonetheless, models of equivalent or higher predictivity were generally obtained within less simulation time using the multi-fidelity approach (increasing Q2 coefficient).

When approximating directly the objective function by a meta-model, the efficiency of the multi-fidelity approach compared to simple kriging turned out to strongly depend on the data included into the objective function. For some of them, the relationship between the two levels of resolution is far from linear. This may explain that the multi-fidelity approach is not efficient. However, very good results were obtained by computing the objective function from the meta-models built to approximate the dynamic responses.

To conclude, the results obtained so far with multi-fidelity meta-modeling on reservoir models are promising. Further tests must be performed on more complex cases. Another perspective concerns the design of experiments with the use of iterative approaches to improve the meta-models.

References

[1] F. Douarche, S. Da Veiga, M. Feraille, G. Enchry, S. Touzani, and R. Barsalou. Sensitivity analysis and optimization of surfactant-polymer flooding under uncertainties. *Oil & Gas Science and Technology*, 2014.

[2] M. Feraille and A. Marrel. Prediction under uncertainty on a mature field. *OGST*, 67(2):193–206, 2012.

[3] F. J. T. Floris, M. D. Bush, M. Cuypers, F. Roggero, and A-R. Syversveen. Methods for quantifying the uncertainty of production forecasts: a comparative study. *Petroleum Geoscience*, 7(S):S87–S96, 2001.

[4] MC Kennedy and A O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[5] L. Le Gratiet. Package "muficokriging" reference manual. *http://cran.r-project.org/web/packages/MuFiCokriging/index.html*, 2013.

Short biography – Arthur Thenon holds a Master in Earth Science from Strasbourg University and an Engineering degree in geophysics from the Ecole et Observatoire des Sciences de la Terre (EOST) also in Strasbourg. His PhD thesis, funded by IFPEN, is to study the potential of multi-fidelity meta-modeling applied to reservoir engineering issues.

Island particle algorithms and their application to rare event estimation

C. Vergé
Ecole Polytechnique & ONERA

Supervisor(s): Prof. Pierre Del Moral (UNSW, Sydney), Prof. Eric Moulines (Telecom ParisTech, Paris) and Dr. Jérôme Morio (ONERA, Toulouse)

Ph.D. expected duration: 2012-2015

Adress: ONERA, Chemin de la Hunière et des Joncherettes, BP 80100, FR-91123 PALAISEAU CEDEX Palaiseau

Email: christelle.verge@onera.fr

Abstract:

Interacting particle systems, a.k.a. particle filter or Sequential Monte Carlo (SMC) methods are known to efficiently sample from sequences of complex distributions in a wide range of applications, including rare event analysis, non-linear filtering, hidden Markov chain parameter estimation, signal processing, financial mathematics (see [3], [2] and the references therein). These algorithms evolve, recursively and randomly in time, a sample of random draws, called *particles*, with associated importance weights. The particle cloud is updated through *selection* and *mutation* operations, where the former duplicates or eliminates, through resampling, particles with large or small importance weights, respectively, while the latter disseminates randomly the particles over the state space and updates accordingly the importance weights for further selection. SMC methods are computationally intensive, as the estimation precision depends upon the particle swarm size, which may be critical in online applications where only a limited computational power is at hand. It is natural to study techniques allowing to reduce the particle swarm size, while ensuring good estimates.

Since the particle interaction due to the selection step, running SMC methods in parallel on multicore processors is not straightforward. Considerable effort has been devoted in the past decade to the design of schemes for the parallel implementation of particle filters, from the totally heuristic to the mathematically well-principled approaches. The natural idea we develop in [5], is to parallelize the algorithm by, instead of considering a single batch of N particles, simply dividing the particle population into N_1 batches of each N_2 particles, also called individuals (i.e., $N = N_1 N_2$), where each batch is referred to as an *island*). In this framework, each island evolves according to the standard SMC scheme subjecting alternately the subpopulation to selection and mutation. Unfortunately, the division of the particle population introduces additional bias which may be of note for moderate island sizes N_2. Thus, we proposed to reduce this bias by performing additional selection also on the *island level* by resampling multinomially the islands according to probabilities proportional to the weight averages over the different subpopulations. Selection on the island level may be performed systematically, as in the *double bootstrap* (B^2) *algorithm* or may be activated adaptively by some criterion measuring the skewness of the island weights (like for Effective Sample Size method). The latter approach will be referred to by us as the *double bootstrap algorithm with adaptive selection on the island level* (B^2ASIL) in [4]. At the end of the day, a sequence of Monte Carlo estimators is obtained by weighing up, using the island weights, the self-normalized empirical measures associated with the different particle islands. But, island interaction prevents island parallelization. So we defined a criterion to determine when island interaction is needed. We chose to base our criterion on mean squared error. See [5] for details.

Convergence, asymptotic and non asymptotic properties of SMC methods have been well studied over the past two decades, see [1] and the references therein. In [4], we present some novel convergence results for island particle models introduced in [5]. In particular we establish a central

limit theorem (CLT)—as the number of islands and the common size of the islands tend jointly to infinity—of the B^2ASIL algorithm. For this purpose we introduce a notion of *archipelagos of weighted islands* that generalizes the particle models studied in [5] and consider three kinds of convergence properties of such archipelagos, namely *consistency* (convergence in probability), *asymptotic normality* (convergence in distribution in terms of a CLT with rate $\sqrt{N}$), and *large deviation* (an exponential inequality of Hoeffing-type that holds uniformly over all islands). The analysis of these properties is challenging due to the strong dependence among the particles through the different operations. So, we perform single-step analyses of three kinds of operations on archipelagos, namely *selection on the island level*, *selection on the individual level* and *mutation*, and we find conditions under which the previous set of convergence properties is preserved by these operations on archipelagos. This theory allows arbitrary compositions of these operations to be straightforwardly analysed, providing a very flexible framework covering the B^2 algorithm as a special case. In the proposed proofs, which rely on limit theorems for triangular arrays, the working process is highly inductive. We also establish the long-term numerical stability of the B^2 algorithm by bounding its asymptotic variance under weak and easily checked assumptions that are typically satisfied for models with non-compact state space.

A potential application of island particle models concerns safety and reliability. Indeed, it is not just evaluating a risk or a probability but estimating the law of random phenomena that leads to critic events. Some parameters of the model or density parameters of input random variables in the system, may be fixed by an experimenter. From a risk analysis point of view, it is interesting to determine the impact of such tuning of parameters on the realisation of some critic event. We developed an algorithm which belongs to the island particle models, referred to as *interacting Particle Markov Chain Monte Carlo (i-PMCMC) algorithm* in [6]. This algorithm samples from the law of parameters of a system conditionally on a rare event. I-PMCMC algorithm consists in running an SMC with N_1 realisations of parameters, and for each of them, running an other SMC with N_2 particles in order to approximate the importance weights (which are often not computable) by unbiased estimators. The interesting result is that the convergence of this algorithm has been established as soon as N_1 gets large, for any value of N_2. We checked this convergence on a test case and we applied this algorithm from the estimation of launch vehicle booster fallout zone.

References

[1] P. Del Moral. *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications*. Springer-Verlag, New York, 2004.

[2] Pierre Del Moral and Christelle Vergé. *Modèles et méthodes stochastiques : une introduction avec applications*, volume 75. Springer Series : Maths & Applications, SMAI, 2014. DOI : 10.1007/978-3-642-54616-7.

[3] A. Doucet, N. De Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. New York, 2001.

[4] C. Vergé, P. Del Moral, E. Moulines, and J. Olsson. Convergence properties of weighted archipelagos with application to the double bootstrap algorithm. Preprint, 2014.

[5] C. Vergé, C. Dubarry, P. Del Moral, and E. Moulines. On parallel implementation of Sequential Monte Carlo methods: the island particle model. *Statistics and Computing*, 23, 2013. DOI : 10.1007/s11222-013-9429-x.

[6] C Vergé, J Morio, and P Del Moral. An island Particle Markov Chain Monte Carlo algorithm for safety analysis. Preprint.

Short biography – Christelle Vergé passed the Agregation of Mathematics in 2011. Then, she graduated from Bordeaux University with an MSc in statistics and stochastic models. Since 2012, she is a PhD student in applied mathematics at ONERA & Ecole Polytechnique, under supervision of P. Del Moral, E. Moulines and J. Morio. Her PhD funding comes from ONERA and CNES. She is interested in parallelization of Sequential Monte Carlo methods and rare event simulations.

Effects of uncertain column alignment in progressive collapse analysis of steel frame structures

Hugues Vincent
Columbia University in the city of New York

Supervisor(s): George Deodatis (Columbia University), Pierre Jehel (CentraleSupélec, MascotNum member) and Simos Gerasimidis (Columbia University)

Ph.D. expected duration: 2015-2018

Address: Department of Civil Engineering and Engineering Mechanics, Columbia University, 630 SW Mudd, 500 West 120th Street, New York, NY 10027, U.S.A.

Email: hugues.vincent@ens-cachan.fr

Abstract:

Progressive collapse can be defined as the ability of a structure to carry loads as a sudden local structural failure phenomenon occurs. This type of phenomenon is to be related to extreme loading such as blast, hurricanes or earthquakes. The interest in the field of progressive collapse increased after recent terrorism attacks (the September 11 attack in New York in 2001) and devastating natural disasters. Assessing the stability of structures in case of extreme loading is not trivial and requires challenging developments to be taken up.

In progressive collapse analysis of frame buildings in blast loading conditions, inelastic buckling of columns appears to be one of the principal instability failure modes [3]. A refined local behavior model including material and geometrical nonlinearities needs to be accounted for to perform such analysis so that the effects of local buckling can be grasped. Moreover, buckling analyses have shown to be very sensitive to small variations in the model description (behavior law, material characteristics, structural geometry, etc.) resulting in possibly large discrepancies in final results and more particularly in variations of the buckling capacity of vertical structural components [5]. Consequently, in order to take reliable decisions from numerical simulations, it is important to be both extremely realistic in the construction of the numerical model and to take into account all the uncertainties (at least the most influencing ones) to be able to assess structural safety. This latter aspect is the center of interest in this study.

Improvements in computational mechanics and numerical resolution allows one to solve Civil Engineering problems in a probabilistic framework that is suitable for dealing with uncertainties. Nevertheless, dealing with uncertainties in the resolution of full-scale structural problems can still be too costly to be within reach. Indeed, probabilistic methods (Monte Carlo methods) requires numerous simulations to be run (usually thousands or tens of thousands) to account for uncertainties such as for instance structural geometry parameters. An interesting study on a whole reinforced concrete structure has highlighted these difficulties to combine numerical simulation on an entire building and refined finite element model with probabilities [4]. Whereas meshing with shell elements, as for instance in [5], would provide accurate representation of local buckling effects that ultimately affect progressive collapse analysis, these latter elements require too much computational resources to be compatible with studies on entire buildings. Thus, efficient sampling techniques as well as efficient computational procedures need to be implemented to account for the effect of uncertainties in the progressive collapse analysis of actual buildings.

In the work presented here, it is investigated how uncertainty in the column alignment propagates to the collapse load of a steel frame building. To that purpose, an uncertain geometry first has to

be parameterized and generated. Columns and beams are supposed to be perfectly straight but not necessarily perfectly vertical or horizontal. Let $\mathbf{X}^n$ denote the position of the $n \in [1, .., N]$ beam-to-column connexions in the perfect configuration (all columns are perfectly vertical and all beams perfectly horizontal). Then, random spatial horizontal perturbations (defects) $\boldsymbol{\delta}(\mathbf{X})$ are generated so that an imperfect configuration is defined by the points (nodes) $\mathbf{x}^n = \mathbf{X}^n + \boldsymbol{\delta}(\mathbf{X}^n)$. The components of $\boldsymbol{\delta}(\mathbf{X})$ are built as realizations of Gaussian homogeneous random field generated according to the Spectral Representation Method [6]. This in particular enables investigating the effects of spatial correlation for the alignment defects [1].

Loading is twofold. On the one hand, service vertical loading forces $\mathbf{Q}(t)$ are applied and, on the other hand, extreme loading is applied either as horizontal forces $\mathbf{H}(t)$ in the case of earthquake (pushover analysis) or as a column removal in the case of blast, then following the alternate load path philosophy commonly adopted [2]. The analyses are performed with the finite element computer program ABAQUS. Quadratic Timoschenko beam elements are used, with geometrical nonlinearities but that omits local buckling to limit computational cost. Depending on the extreme loading considered, either $\mathbf{Q}(t)$ or $\mathbf{H}(t)$ is incrementally increased until structural collapse is detected. A sensitivity analysis is performed to study the impact of uncertainty in column alignment on the collapse load.

Future work will be oriented toward introducing other sources of uncertainty along with more refined structural models that are capable of grasping local buckling effects. This may lead to the development of surrogate models to tackle the issue of the large computational resources required.

References

[1] P. Bocchini, D.M. Frangopol, and G. Deodatis. A random field based technique for the efficiency enhancement of bridge network life-cycle analysis under uncertainty. *Engineering Structures*, 33:3208–3217, 2011.

[2] DoD. Unified facilities criteria (UFC), design of buildings to resist progressive collapse. Technical report, Department of Defense, Washington, DC, USA, 2009.

[3] S. Gerasimidis, G. Deodatis, T. Kontoroupi, and M. Ettouney. Loss-of-stability induced progressive collapse modes in 3D steel moment frames. *Structure and Infrastructure Engineering*, 2014. DOI: 10.1080/15732479.2014.885063.

[4] D. Kelliher and K. Sutton-Swaby. Stochastic representation of blast load damage in a reinforced concrete building. *Structural Safety*, 34:407–417, 2012.

[5] V. Papadopoulos, G. Soimiris, and M. Papadrakakis. Buckling analysis of i-section portal frames with stochastic imperfections. *Engineering Structures*, 47:54–66, 2013.

[6] M. Shinozuka and G. Deodatis. Simulation of multi-dimensional gaussian stochastic fields by spectral representation. *Applied Mechanics Review (ASME)*, 49:29–53, 1996.

Short biography – I graduated from ENS Cachan with a Master of science in June 2014. I am a research scholar at Columbia University with Professor George Deodatis for one year. My research work is oriented toward quantifying the effects of uncertainties in numerical progressive collapse analysis of buildings. This research work is supported by ENS Cachan, Mécénat Besnard de Quelen, which awarded me a scholarship, and by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme.

Point Process-based estimation of k^{th}-order moment

C. Walter
Université Paris Diderot - Paris 7
Commissariat à l'Énergie Atomique et aux Énergies Alternatives

Supervisor(s): Prof. Josselin Garnier (Université Paris Diderot - Paris 7) and Gilles Defaux (CEA)

Ph.D. expected duration: 2013-2016

Address: CEA, DAM, DIF, F-91297 Arpajon, France

Email: clement.walter@cea.fr

Abstract: The estimation of k^{th}-order moment of a real-valued random variable is often done by the mean of Monte Carlo method [9]. It finds its mathematical support combining the Strong Law of Large Numbers (SLLN) and the Central Limit Theorem (CLT). The first one gives an almost surely convergence to the sough moment while the second one gives a convergence in law of the estimator. One of the main drawbacks of this method is that it requires the existence of a moment of order $2k$ to have a finite variance. This limitation may be critical when dealing with heavy-tailed distributions, like Pareto or Pareto type distributions.

In this scope, variance-reduction techniques such as Importance Sampling (IS) can circumvent this limitation. Basically, IS changes the measure of the random variable of interest to lower the variance, and possibly to make it finite. However, it is intrusive as it requires to generate samples according to an auxiliary distribution and it is not always possible to have access to the optimal change of measure. In this latter case, the variance of the modified estimator can even be worse (see [13] or [4] for further details on Importance Sampling).

Other work has been devoted to order statistics, tail index estimation and trimmed sums [3, 14, 5, 2]. While producing finite-variance estimators that are unbiased or with a controlled bias, they always assume parametric forms for the distribution functions and give only limited results; see [1] for a comprehensive overview of tail index estimation, and [11, 7, 10, 6] for references on mean estimation for heavy-tailed random variables.

Using recent results on Point Processes related to a real-valued random variable with continuous *cdf* [15] we first introduce a new general estimator for k^{th}-order moment. This estimator is unbiased with finite variance as soon as a moment of order k' exists for some $k' > k$, and is always better than the usual Monte Carlo method in terms of variance. Unfortunately, it requires to simulate an infinite sum. To handle this problem we propose to make use of new results on exact path simulation [8, 12] to develop a randomised unbiased estimator which has also a finite variance as soon as a moment of order $k' > k$ exists. This randomised estimator supports also a CLT.

Considering its implementation, we derive optimal parameters depending on the distribution of the random variable of interest. We further show that there is not a huge gain in term of variance between this optimal implementation and a suboptimal one with prescribed parameters. Hence, we have produced a new versatile estimator for k^{th}-order moment which only requires to be able to generate samples according to the distribution of the variable of interest (for instance the output of a complex numerical code with random inputs).

References

[1] Jan Beirlant, Frederico Caeiro, and M Ivette Gomes. An overview and open research topics in statistics of univariate extremes. *REVSTAT-Statistical Journal*, 10(1):1–31, 2012.

[2] István Berkes and Lajos Horváth. The central limit theorem for sums of trimmed variables with heavy tails. *Stochastic Processes and their Applications*, 122(2):449–465, 2012.

[3] Peter J Bickel et al. On some robust estimates of location. *The Annals of Mathematical Statistics*, 36(3):847–858, 1965.

[4] Peter W Glynn and Donald L Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.

[5] Philip S Griffin. Asymptotic normality of winsorized means. *Stochastic processes and their applications*, 29(1):107–127, 1988.

[6] Jonathan B Hill. Robust estimation for average treatment effects. *Available at SSRN 2260573*, 2013.

[7] Joachim Johansson. Estimating the mean of heavy-tailed distributions. *Extremes*, 6(2):91–109, 2003.

[8] Dan McLeish. A general method for debiasing a monte carlo estimator. *Monte Carlo Methods and Applications*, 2011.

[9] Nicholas Metropolis and Stanislaw Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[10] Abdelhakim Necir, Abdelaziz Rassoul, and Ričardas Zitikis. Estimating the conditional tail expectation in the case of heavy-tailed losses. *Journal of Probability and Statistics*, 2010, 2010.

[11] Liang Peng. Estimating the mean of a heavy tailed distribution. *Statistics & Probability Letters*, 52(3):255–264, 2001.

[12] Chang-han Rhee and Peter W Glynn. Unbiased estimation with square root convergence for sde models. 2013.

[13] Christian P Robert and George Casella. *Monte Carlo statistical methods*. Springer, 2004.

[14] Stephen M. Stigler. The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1(3):pp. 472–477, 1973.

[15] C. Walter. Moving Particles: a parallel optimal Multilevel Splitting method with application in quantiles estimation and meta-model based algorithms. *ArXiv 1405.2800 e-prints*, May 2014.

Short biography – I am 25 years old. I graduated from Mines ParisTech in 2013. At Mines Paristech I specialised in geostatistics and used it for natural resources evaluation. I discovered the use of Kriging in computer sciences at CEA where I did my final internship during summer 2013. Then I pursued with a PhD in rare event simulation, which I started in November 2013.