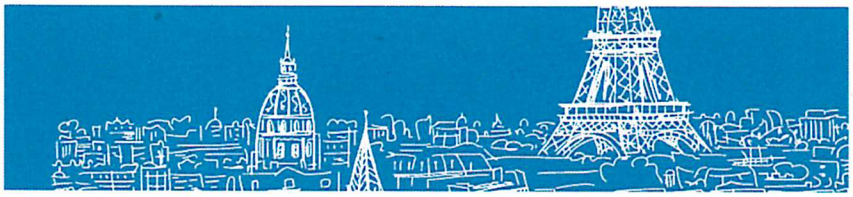
MASCOT-NUM 2019



18-20 March 2019 - IFPEN, Rueil-Malmaison (near Paris)

Abstract Volume

IFP Energies nouvelles

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

Oral sessions

MASCOT-NUM 2019

PRIEUR Clémentine
Grenoble Alpes University, Jean Kuntzmann Lab., Inria project/team AIRSEA

Abstract

Dimension reduction of the input parameter space for potentially vector-valued functions

Many problems that arise in uncertainty quantification, e.g., integrating or approximating multivariate functions, suffer from the curse of dimensionality. The cost of computing a sufficiently accurate approximation grows indeed dramatically with the dimension of the input parameter space. It thus seems important to identify and exploit some notion of low-dimensional structure as, e.g., the intrinsic dimension of the model. A function varying primarily along a few directions of the input parameter space is said of low intrinsic dimension. In that setting, algorithms for quantifying uncertainty focusing on these important directions are expected to reduce the overall cost. A common approach to reducing a function's input dimension is the truncated Karhunen-Loève decomposition [1], which exploits the correlation structure of the function's input space. In the present talk, we propose to exploit not only input correlations but also the structure of the input-output map itself. We will first focus the presentation on approaches based on global sensitivity analysis. The main drawback of global sensitivity analysis is the cost required to estimate sensitivity indices such as Sobol' indices [2]. It is the main reason why we turn to the notion of active subspaces [3, 4] defined as eigenspaces of the average outer product of the function's gradient with itself. They capture the directions along which the function varies the most, in the sense of its output responding most strongly to input perturbations, in expectation over the input measure. In particular, we will present recent results stated in [5] dealing with the framework of multivariate vector-valued functions.

References

[1] C. Schwab, R. A. Todor, Karhunen–Loève approximation of random fields by generalized fast multipole methods, Journal of Computational Physics 217 (1) (2006) 100–122.

[2] I. Sobol, Sensitivity estimates for non linear mathematical models, Mathematical Modelling and Computational Experiments 1 (1993) 407–414.

[3] P. G. Constantine, E. Dow, Q.Wang, Active subspace methods in theory and practice: Applications to kriging surfaces, SIAM Journal on Scientific Computing 36 (4) (2014) A1500–A1524.

[4] P. G. Constantine, Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies, Society for Industrial and Applied Mathematics, Philadelphia, 2015.

[5] O. Zahm, P. Constantine, C. Prieur, Y. Marzouk, Gradient-based dimension reduction of multivariate vector-valued functions (2018). URL https://hal.inria.fr/hal-01701425

MascotNum2019 conference - Estimation of Borgonovo's moment independent importance measures

DERENNES PIERRE
Université Paul Sabatier (Toulouse).

Supervisor(s): Prof. J. Morio (ONERA Toulouse) and Prof. F. Simatos (ISAE Toulouse).

Ph.D. expected duration: Sep. 2016 - Jul. 2019

Adress: ONERA - The French Aerospace Lab; 2 Avenue Edouard Belin, 31000 Toulouse.

Email: pierre.derennes@onera.fr

Abstract:

In diverse disciplines, complex systems modeling is often achieved by considering a black-box model for which the observation is expressed as a deterministic function of external parameters representing some physical variables. These basic variables are usually assumed random in order to take phenomenological uncertainties into account. Then, *global sensitivity analysis* (GSA) techniques play a crucial role in the handling of these uncertainties and in the comprehension of the system behavior.

Variance-based sensitivity indices are one of the most widely used GSA measures. They are based on Sobol's indices which express the share of variance of the output that is due to a given input or input combination. However, by definition they only study the impact on the second-order moment of the output which is a restricted representation of the whole output distribution. Moment independent importance measures have been proposed by E. Borgonovo [1] in order to alleviate this drawback. Throughout we consider a general input-output model $Y = \mathcal{M}(X_1, ...X_d)$ where the scalar output Y depends on a d-dimensional real valued random variable $\mathbf{X} = (X_1, ..., X_d)$ through a deterministic function $\mathcal{M}$. We assume that for every $I \subset \{1, \dots, d\}$ a strict subset, the pair $(\mathbf{X}_I, Y)$ is absolutely continuous. The idea of Borgonovo's GSA approach is to measure how fixing $\mathbf{X}_I$ at a value $\mathbf{x}_I$ modifies the entire distribution of the output Y. This modification is quantified by the *shift* $s(\mathbf{x}_I)$ defined from the L_1-norm between the output probability density function (PDF) f_Y and the conditional output PDF $f_Y^{\mathbf{X}_I=\mathbf{x}_I}$:

$$s(\mathbf{x}_I) = \frac{1}{2} \left\| f_Y - f_Y^{\mathbf{X}_I=\mathbf{x}_I} \right\|_{L^1(\mathbb{R})} = \frac{1}{2} \int \left| f_Y(y) - f_Y^{\mathbf{X}_I=\mathbf{x}_I}(y) \right| dy \ . \tag{1}$$

So as to consider the whole range of values the random variable $\mathbf{X}_I$ can take into account, the sensibility of the output Y with respect to the input $\mathbf{X}_I$ is defined by averaging the shift over $\mathbf{X}_I$:

$$\delta_I := \mathbb{E}\left[s(\mathbf{X}_I)\right] \ . \tag{2}$$

Estimating Borgonovo's indices is a challenging task because of the unknown unconditional and conditional PDFs f_Y and $f_Y^{\mathbf{X}_I=\mathbf{x}_I}$ that intervene in a convoluted way (i.e., through an L_1-norm) in their definitions (1) and (2). The estimation of first order indices δ_i has been the subject of extensive investigation in several works, see for instance [4][6][7].

In [6], it is shown that Borgonovo's indices can be reinterpreted as a dependence measure:

$$\delta_I = \frac{1}{2} \left\| f_{\mathbf{X}_I} f_Y - f_{\mathbf{X}_I, Y} \right\|_{L_1(\mathbb{R}^2)} \ . \tag{3}$$

The first contribution of this work [3] consists in introducing a new estimation scheme of δ_i measures from the definition (3). The proposed method combines importance sampling and the

Gaussian kernel estimation of the output PDF f_Y and joint PDF $f_{X_i,Y}$ and enables the estimation of all the first order indices δ_i from one data set. Furthermore, some theoretical properties of the proposed estimator, and in particular its consistency, are derived. However, some limitations exist due to the use of Gaussian kernel estimation which may be inefficient in the case of bounded support or heavy tailed distributions.

The second contribution is to define a new estimation approach avoiding these drawbacks. The starting point is the observation made by [5] that δ_i indices can be expressed in terms of *copula density*:

$$\delta_i = \frac{1}{2} \int_{[0,1]^2} |c_i(u,v) - 1| \, dudv \; . \tag{4}$$

where c_i denotes the copula density of the pair (X_i, Y), i.e, the PDF of $(F_{X_i}(X_i), F_Y(Y))$. In [2], it is proposed to estimate the double integral in Eq.(4) from a simple Monte Carlo estimation coupled to the maximum entropy estimation of the bivariate density copula c_i with *fractional moments* constraints. Some promising results are obtained on different test cases [2].

The application of this work to reliability analysis framework is an interesting perspective. Practitioners often seek to estimate a so-called failure probability associated to an unsafe and undesired state of the considered system. This uncertainty propagation phase may be completed by considering the δ-sensitivity measures of the system conditioned on this failure state. Another perspective is the estimation of higher order δ-indices which relies on being able to estimate high-dimensional PDFs. The difficulty of this problem lies in the *curse of dimensionality*. Indeed the computation burden required for sufficiently accurate estimates obtained with classical methods like kernel estimation and maximum entropy estimation excessively grows when the dimension increases.

References

[1] Emanuele Borgonovo. A new uncertainty importance measure. *Reliability Engineering & System Safety*, 92(6):771–784, 2007.

[2] Pierre Derennes, Jérôme Morio, and Florian Simatos. Estimation of the moment independent importance sampling measures using copula and maximum entropy framework. *Winter Simulation Conference. At: Gothenburg Sweden*, 2018.

[3] Pierre Derennes, Jérôme Morio, and Florian Simatos. A nonparametric importance sampling estimator for moment independent importance measures. *Reliability Engineering & System Safety*, (In press), 2018.

[4] Qiao Liu and Toshimitsu Homma. A new computational method of a moment-independent uncertainty importance measure. *Reliability Engineering & System Safety*, 94(7):1205–1211, 2009.

[5] Pengfei Wei, Zhenzhou Lu, and Jingwen Song. Moment-independent sensitivity analysis using copula. *Risk Analysis*, 34(2):210–222, 2014.

[6] Pengfei Wei, Zhenzhou Lu, and Xiukai Yuan. Monte Carlo simulation for moment-independent sensitivity analysis. *Reliability Engineering & System Safety*, 110:60–67, 2013.

[7] Leigang Zhang, Zhenzhou Lu, Lei Cheng, and Chongqing Fan. A new method for evaluating borgonovo moment-independent importance measure with its application in an aircraft structure. *Reliability Engineering & System Safety*, 132:163–175, 2014.

Short biography – Pierre Derennes is a second year PhD student of applied mathematics funded by Université Paul Sabatier in Toulouse. His home lab is ONERA (Toulouse). He obtained his master's degree at Université Claude-Bernard in Lyon.

Bayesian optimization in effective dimensions via kernel-based sensitivity indices

A. SPAGNOL
Ecole des Mines de Saint-Etienne / Safran Tech

Supervisor(s): R. Le Riche (CNRS / Ecole des Mines de Saint-Etienne) and S. Da Veiga (Safran Tech)

Ph.D. expected duration: Feb. 2017 - Jan. 2020

Adress: Safran Tech, 1 Rue Genevieve Aube, 78114 Magny-les-Hameaux

Email: adrien.spagnol@safrangroup.com

Abstract:

Optimization of high dimensional functions under constraints and reliability assessment are key engineering problems, but they often come at a prohibitive cost since they usually involve a complex or expensive computer code. To overcome this limitation, analysts frequently rely on a preliminary dimension reduction by identifying which parameters drive the most the function variations: non-influential variables are set to a fixed value and optimization or reliability procedures are carried out with the remaining, significant, variables. Yet, the classical influence measures, which are meaningful for regression problems, do not account for the specific structure of optimization or reliability problems and can even lead to inaccurate solutions.

In this work, we describe a recent sensitivity index defined through a kernel-based dependency measure, the Hilbert Schmidt Independence Criterion [2]. This HSIC measure is designed to characterize whether a design variable matters to reach low values of the objective function and to satisfy the constraints. Such sensitivity criterion can readily be extended to reliability levels.

Finally, inspired by recent works in Gaussian Process-based optimization, where the authors only optimize on a randomly drawn subset of relevant variables at each iteration [1], we use this sensitivity measure to guide the selection. Our method either picks variables in a probabilistic manner where the subset of effective variables is drawn at random with probabilities equal to the normalized HSIC measures, or in a deterministic one keeping only the variables whose normalized HSIC measure is above a given threshold. We also provide different strategies to deal with the negligible inputs and apply our method on several examples from optimization benchmarks, as in Figure 1, to demonstrate how clever variable selection can efficiently improve the optimization.

References

[1] Cheng Li, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh, and Alistair Shilton. High dimensional bayesian optimization using dropout. *preprint arXiv:1802.05400*, 2018.

[2] Adrien Spagnol, Rodolphe Le Riche, and Sébastien Da Veiga. Global sensitivity analysis for optimization with variable selection. *arXiv preprint arXiv:1811.04646*, 2018.

Short biography – Adrien Spagnol is a second-year PhD student in applied mathematics at Safran Tech, in collaboration with the Ecole des Mines de Saint-Etienne. He received a master's degree in structural and mechanical engineering from the French Institute of Mechanics in Clermont-Ferrand (France).

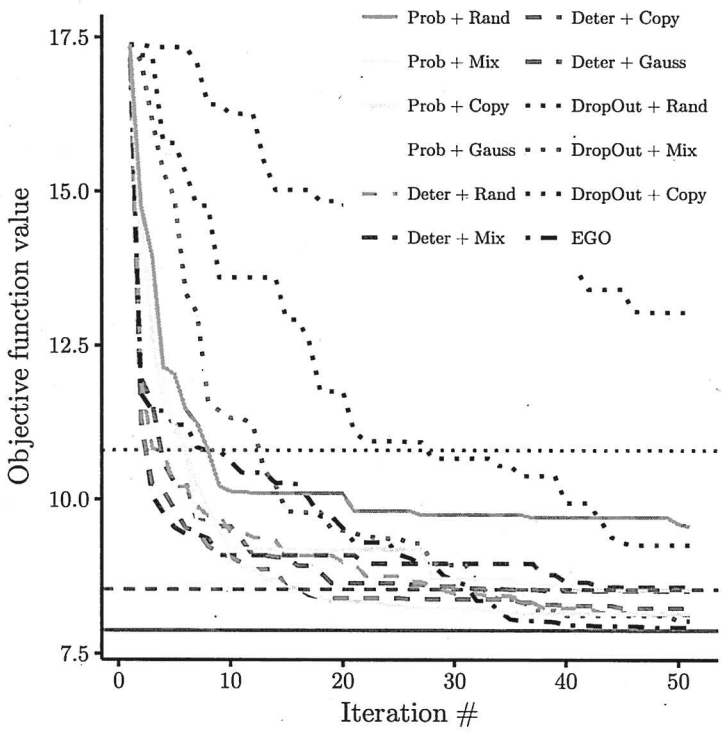


Figure 1: Median results of the different algorithms for the Borehole function. The red lines correspond to the easy, medium and hard goals (from top to bottom) for this test case, which are defined as the 90%, 50% and 10% quantiles of the final results of all algorithms.

Gaussian process regression models under linear inequality conditions

A. F. López-Lopera
Mines Saint-Étienne (EMSE), France.

Supervisor(s): O. Roustant (EMSE, France), F. Bachoc (Institut de Mathématiques de Toulouse, France) and N. Durrande (EMSE, France; PROWLER.io, Cambridge, UK)

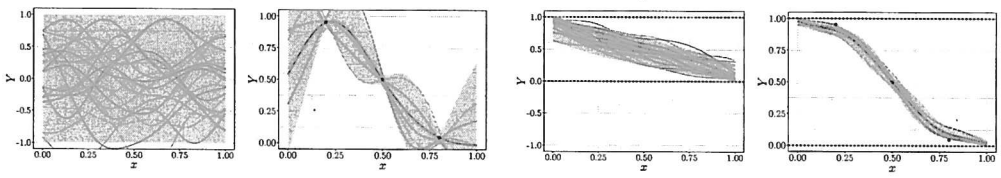
Ph.D. expected duration: Oct. 2016 - Sep. 2019

Adress: Institut Henri FAYOL – 158 cours Fauriel F-42023 Saint-Étienne, France.

Email: andres-felipe.lopez@emse.fr

Abstract:

Taking into account inequality constraints (e.g. boundedness, monotonicity, convexity) into Gaussian process (GP) models can lead to more realistic predictions guided by the physics of data [6, 4]. Figure 1 compares two models that either ignore or take into account both boundedness (i.e. $0 \leq y(x) \leq 1$, for $x \in [0, 1]$) and monotonicity constraints (i.e. $y(x) \geq y(x')$, if $x \geq x'$).



(a) unconstrained GP (b) constrained GP

Figure 1: GP regression models under (a) no constraints and (b) boundedness and monotonicity constraints. Each panel shows: (left) samples from the different types of Gaussian priors, and (right) the resulting GP regression model conditioned on three observations (dots).

We aim at investigating a GP framework that can account for inequality constraints. Our main contributions are threefold.

First, building on the approach proposed in [6], we introduced in [4] a full Gaussian-based framework to satisfy a set of linear inequality constraints. The benefit of using the finite-dimensional representation of [6] leads to satisfy the inequalities everywhere in the input space. Furthermore, it was proved in [2] that the resulting posterior mode is the optimal constrained interpolation function in the reproducing kernel Hilbert space. Due to the truncated Gaussianity of the posterior, its distribution can be approximated via Monte Carlo or Markov chain Monte Carlo. We investigated several samplers in examples on both synthetic and real-world data, under different types of constraints. We found that the Hamiltonian Monte Carlo (HMC)-based sampler from [7] achieves the best trade-off between running time and effective sample rates.

Despite the promising results in [4], our experiments were limited up to 2D problems due to the tensor structure of our framework. This brings us to our second contribution, where various alternatives have been explored for going to higher dimensions and for a high number of observations. In the first direction, we introduced noise for the relaxation of the interpolation constraints. This also relaxed the constraints of the HMC sampler improving its efficiency. As a result, we were now able to use our framework in 5D spaces [3]. Moreover, since the computational complexity here depends on the number of basis functions rather than the observations, we can

handled thousands of observations. In a second direction, we considered specific assumptions on the target function that are suitable in high dimensions. In particular, we adapted our framework to additive functions, where sampling from the posterior distribution in high dimensions can be achieved through sampling in lower dimensional spaces (e.g. 1D spaces by assuming first-order additivity) [5]. Figure 2 plots a 5D example for a first-order additive target function satisfying different types of inequality constraints per dimension.

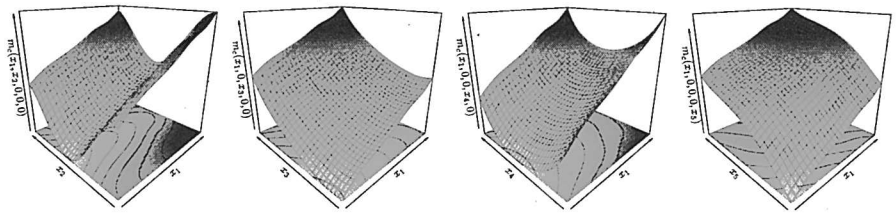


Figure 2: Additive GP regression model for $\mathbf{x} \mapsto 2x_1 + \cos(6x_2) + 2x_3^2 + 4(x_4 - 0.5)^2 + 2\arctan(2x_5)$ for $\mathbf{x} \in [0,1]^5$. The constrained predictive mean is shown satisfying: monotonicity constraints across the first and fifth dimensions, and convexity constraints across the third and fourth dimensions. No constraints were imposed across the second dimension.

Third, we considered the problem of estimating the covariance parameter under inequality constraints. We studied the properties of both unconstrained and constrained maximum likelihood (ML) estimators. Under fixed-domain asymptotics, we showed that, loosely speaking, any consistency result for the (unconstrained) ML is preserved for the constrained ML when adding boundedness, monotonicity and convexity conditions [4]. We also showed that the constrained ML estimator (cMLE), conditionally to the fact that the GP satisfies those constraints, has the same asymptotic distribution as the unconditional MLE [1].

References

[1] F. Bachoc, A. Lagnoux, and A. F. López-Lopera. Maximum likelihood estimation for Gaussian processes under inequality constraints. *ArXiv e-prints*, April 2018.

[2] X. Bay, L. Grammont, and H. Maatouk. Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation. *Electron. J. Stat.*, 10(1):1580–1595, 2016.

[3] A. F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant. Approximating Gaussian process emulators with inequality constraints and noisy observations (submitted). 2018.

[4] A. F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant. Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1224–1255, 2018.

[5] A. F. López-Lopera, N. Durrande, F. Bachoc, and O. Roustant. Additive Gaussian processes under inequality constraints (in preparation). 2019.

[6] H. Maatouk and X. Bay. Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582, 2017.

[7] Ari Pakman and Liam Paninski. Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, 2014.

Short biography – A. F. López-Lopera received the BEng and MEng degrees in electrical engineering from the *Universidad Tecnológica de Pereira*, Colombia. Currently, he is a PhD student in applied mathematics at EMSE, France. In the thesis, *Metamodelling under inequality constraints*, GP-based models are explored in two main directions: the estimation under inequality constraints, and the extension to higher dimensions. His PhD is funded by the chair in applied mathematics OQUAIDO.

Dimension Reduction for the Bayesian Optimization of Shapes

D. GAUDRIE
Mines de Saint-Étienne - Groupe PSA

Supervisor(s): Dr. R. Le Riche (CNRS, Mines de Saint-Étienne), Dr. V. Picheny (Prowler.io), Dr. B. Enaux (Groupe PSA) and Dr. V. Herbert (Groupe PSA)

Ph.D. expected duration: 2016-2019

Address: Mines de Saint-Étienne, 29 Rue Pierre et Dominique Ponchardier, 42100 Saint-Étienne

Email: david.gaudrie@mpsa.com

Abstract:

Parametric shape optimization aims at minimizing one (or $m \geq 2$ in a multi-objective setting) objective function $f(\mathbf{x})$ where $\mathbf{x} \in X \subset \mathbb{R}^d$ is a d-dimensional vector of CAD parameters. It is common that d is large, $d \gtrsim 50$. Optimization in such a high-dimensional design space is difficult, especially when $f(\cdot)$ is an expensive black-box function and the use of surrogate-based approaches [1] is mandatory. The ratio between the allowed budget of function evaluations ($b \approx$ 100-200) and d is also too small to perform sensitivity analysis prior to selecting $d' << d$ variables.
In this work, we exploit the fact that the computation time of a shape $\Omega_{\mathbf{x}}$ is negligible in comparison with the evaluation time of $f(\mathbf{x})$. Most often, the set of all CAD generated shapes, $\mathbf{\Omega} := \{\Omega_{\mathbf{x}}, \mathbf{x} \in X\}$ can be approximated in a $\delta << d$ -dimensional manifold where it is preferable to build the surrogate model and perform the optimization. To uncover this manifold, we apply PCA to a dataset of designs and test alternative shape representations. We then build Gaussian processes and optimize in the reduced space of eigenshapes. Such approaches have already been considered in part in [3, 4], but we provide a new integrated view of shape reduction and optimization with kernel methods. In the following, the essential elements of our approach are further introduced.

From CAD description to shape eigenbasis. Let $\phi : X \to \Phi$ be a mapping to a high-dimensional space $\Phi \subset \mathbb{R}^D$, $D >> d$. We have compared alternative $\phi : X \to \Phi$ based on their ability to uncover intrinsic dimensions through Principal Component Analysis (PCA). The ϕ studied here are the characteristic function, the signed distance to contours and the contour discretization. We proceed by uniformly sampling N designs in X. Performing a PCA of that sample in the X space would be useless. However, with a proper choice of ϕ, we have found that a few (δ) eigenshapes allow to accurately describe the sample of CAD shapes through their principal components, $\boldsymbol{\alpha}$ in the eigenbasis $(\mathbf{v}^1, \ldots, \mathbf{v}^D)$.

GP model for shrinked spaces. Instead of building a surrogate for $f(\cdot)$ using $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \in X \subset \mathbb{R}^d$, a GP is fitted to the principal components $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(n)} \in \mathbb{R}^D$. To simultaneously emphasize the δ most important axes without entirely neglecting the $D - \delta$ remaining ones, an additive model between the (δ) active components and the residual coordinates is considered:

$$Y(\boldsymbol{\alpha}) = \mu + Y^a(\boldsymbol{\alpha}_{1:\delta}) + Y^{\overline{a}}(\boldsymbol{\alpha}_{\delta+1:D}) + \varepsilon \ .$$

$Y^a(\cdot) \sim \mathcal{GP}_\delta(0, k^a(\cdot,\cdot))$ is the main-effect GP in a δ-dimensional input space and $Y^{\overline{a}}(\cdot) \sim \mathcal{GP}_{D-\delta}(0, k^{\overline{a}}(\cdot,\cdot))$ is a sparse, isotropic GP. $Y^{\overline{a}}(\cdot)$ lives in a high dimensional space but only requires the estimation of 2 hyperparameters, θ_D and σ_D^2. It aims at taking the less relevant, though existing, effects of the remaining eigenshapes into account.

Optimization in a reduced space. The well known Expected Improvement [2] is then maximized with the full $Y(\cdot)$ to optimize the shape. It is possible to carry out this maximization in the X space thanks to the ϕ mapping, $\mathbf{x}^{(n+1)} = \underset{\mathbf{x}\in X}{\arg\max}\ \mathrm{EI}(\boldsymbol{\alpha}(\mathbf{x}))$. But such an approach does not take advantage of the space reduction beyond the construction of $Y(\cdot)$. We thus propose a redefinition of improvement to carry out the maximization in the smaller space of important eigenshapes, completed by a cheap maximization with regard to the dimensions $\delta+1$ to D, $\boldsymbol{\alpha}^{(n+1)} = \underset{[\boldsymbol{\alpha}_{1:\delta},\boldsymbol{\alpha}_{\delta+1:D}]\in\mathbb{R}^D}{\arg\max}\ \mathrm{EI}([\boldsymbol{\alpha}_{1:\delta},\boldsymbol{\alpha}_{\delta+1:D}])$. The calculation of the pre-image, $\mathbf{x}^{(n+1)} = \underset{\mathbf{x}\in X}{\arg\min}\|\mathbf{V}^\top\Phi(\mathbf{x}) - \boldsymbol{\alpha}^{(n+1)}\|^2$, is finally performed to find the next parametric design to be evaluated by the computer code.

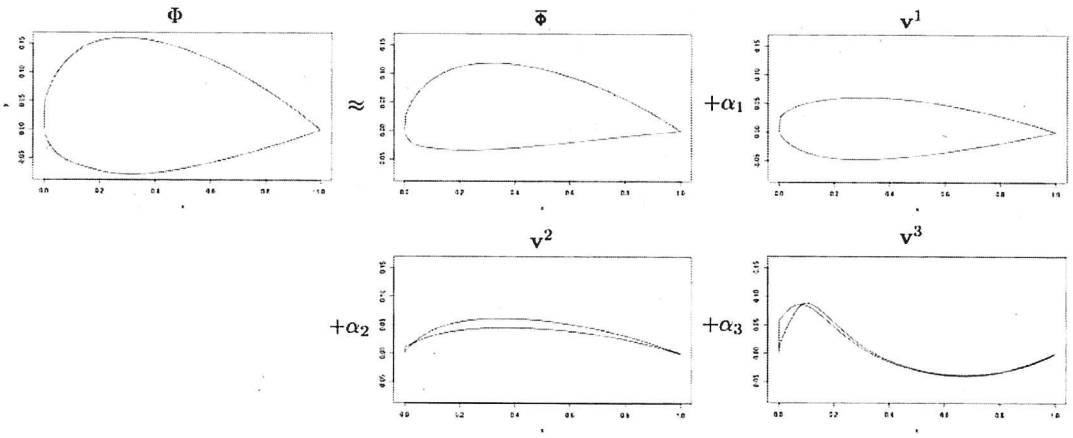


Figure 1: Shape decomposition in its eigenbasis

References

[1] 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.

[2] J Močkus. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer, 1975.

[3] Balaji Raghavan, Guenhael Le Quilliec, Piotr Breitkopf, Alain Rassineux, Jean-Marc Roelandt, and Pierre Villon. Numerical assessment of springback for the deep drawing process by level set interpolation using shape manifolds. *International journal of material forming*, 7(4):487–501, 2014.

[4] Mikkel B Stegmann and David Delgado Gomez. A brief introduction to statistical shape analysis. *Informatics and mathematical modelling, Technical University of Denmark, DTU*, 15(11), 2002.

Short biography – David Gaudrie obtained his engineering degree from INSA Toulouse in Applied Mathematics in 2016. He started a PhD thesis about high dimensional multi-objective optimization in the context of expensive computer codes in November 2016. This thesis in funded by the automotive group PSA (CIFRE convention) in collaboration with the École des Mines de Saint-Étienne.

Stochastic Inversion Under Functional Uncertainties

M.R. El Amri
University of Grenoble Alpes, IFPEN

Supervisor(s): C. Prieur (UGA), C. Helbert (ICJ, ECL), D. Sinoquet (IFPEN), M. Munoz Zuniga (IFPEN), O. Lepreux (IFPEN)

Ph.D. expected duration: 2016 - 2019

Adress: Bâtiment IMAG, 700 Avenue Centrale, 38401 Saint-Martin-d'Hères

Email: mohamed-reda.el-amri@ifpen.fr

Abstract:

In this present work we propose a new efficient method for solving an inversion problem under functional uncertainties. Without loss of generality, the considered numerical simulator takes inputs that can be divided into two sets, the deterministic control variables and the functional random variables. Let $f : \mathbb{X} \times \mathcal{V} \to \mathbb{R}$ denote the output of the simulator where $\mathbb{X} \subset \mathbb{R}^p$ is the search space of the control variables and the randomness of the simulator is derived from a random input defined in a functional space $\mathcal{V}$. We assume that the probability distribution of the functional input is only known through a finite set of realizations $\Xi = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and each evaluation of f involves a time consuming computation. Let $\mathcal{X}_n = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ and $\mathcal{V}_n = \{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ denote the initial design points. The simulator responses at these design points are denoted by $\mathbf{Z} = \{f(\mathbf{x}^1, \mathbf{v}^1), \ldots, f(\mathbf{x}^n, \mathbf{v}^n)\}$.

The objective of this study is to estimate the set

$$\Gamma^* = \{\mathbf{x} \in \mathbb{X} \ , \ g(\mathbf{x}) = \mathbb{E}_{\mathbf{V}}[f(\mathbf{x}, \mathbf{V})] \leq c\}, \tag{1}$$

where $c \in \mathbb{R}$ and $\mathbb{E}_{\mathbf{V}}$ is the expectation with respect to $\mathbf{V}$. The estimation of Γ^* based on computing function g at each grid point of discretized version of $\mathbb{X}$ requires far too many expensive simulations of f. Therefore, statistical methods based on a reduced number of evaluation points are widely used to overcome this latter difficulty by focusing the evaluations on the promising subregion of the control space.

Among statistical models, Gaussian Process (GP) model has received increasing interest in recent years, due to many of its good properties, such as the existence of explicit formula of statistical moments and the easy computation of the uncertainties on predictions. However, in the literature the input variables involved in Gaussian Process models are often univariate or multivariate. The purpose of this talk is first to extend the use of Gaussian Process model to cases where the inputs contain infinite dimensional variables or functional data which are collected as curves. Then we define an infill sampling criterion based on the GP model in order to solve the stochastic inversion problem (1).

After a dimension reduction of the functional space ($\mathbf{V} \in \mathcal{V} \underset{\text{Karhunen-Loève}}{\longrightarrow} \mathbf{U} \in \mathbb{R}^{m_{\text{KL}}}$), a GP model $Z_{(\mathbf{x},\mathbf{u})}$ is built in the joint space of control and uncertain variables $(\mathbf{x}, \mathbf{u})$. Then an averaged GP over the random variables is derived [2], $Y_{(\mathbf{x})} = \mathbb{E}_{\mathbf{U}}[Z_{(\mathbf{x},\mathbf{u})}]$. Therefore the 'induced' Gaussian process approximates the expected response involved in inversion problem (1). Driven by the Gaussian process conditioned on the n observations $Y^n_{(\mathbf{x})}$, the proposed infill strategy consists of two steps: first we choose $\mathbf{x}^{n+1}$ by minimizing the Vorob'ev deviation [1]. Secondly we choose the uncertain point $\mathbf{u}^{n+1}$ that minimizes the variance of the process $Y^n_{(\mathbf{x})}$ evaluated at the point $\mathbf{x}^{n+1}$. By iterating this procedure, we expect a rapid estimation of the set Γ^* with a limited number of calls to the simulator f.

To illustrate the performance of the proposed method we compare it with two approaches. The first one is the nested MC approach, where the expectation with respect to $\mathbf{V}$ is estimated by Monte Carlo simulations using the simulator f, and the inversion is based on a GP model in the control space $\mathbb{X}$. The accuracy level of the MC estimation of the expectation is controlled by the number of calls k to the underlying simulator $f(\mathbf{x}, \mathbf{v})$.

The second method used for comparison is similar to the first one except that the expectation is estimated by a *Greedy Functional Quantizer* of size k. Briefly, the *Functional Quantization* of a functional random variable $\mathbf{V}$ consists of its approximation by a discrete random variable $\hat{\mathbf{V}}_k$ valued in $\Theta_k = \{\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_k\}$ [3]. The "grid" (or codebook) Θ_k minimizes over $(\mathcal{V})^k$ the quantization error induced by replacing $\mathbf{V}$ by $\hat{\mathbf{V}}_k$.

After presenting comparative results on a toy problem, we focus on the results obtained for an automotive industrial test case where the objective is to identify the set of control parameters leading to meet the pollutant emissions standards of a vehicle.

References

[1] Clément Chevalier. *Fast uncertainty reduction strategies relying on Gaussian process models.* PhD thesis, Citeseer, 2013.

[2] Janis Janusevskis and Rodolphe Le Riche. Simultaneous kriging-based estimation and optimization of mean response. *Journal of Global Optimization*, 55(2):313–336, 2013.

[3] Harald Luschgy, Gilles Pagès, and Benedikt Wilbertz. Asymptotically optimal quantization schemes for gaussian processes on hilbert spaces. *ESAIM: Probability and Statistics*, 14:93–116, 2010.

Short biography – I obtained a MSc in Applied mathematics : Computer and Stochastic Methods for Decision in 2015 at Université de Pau et des Pays de l'Adour. I started my PhD with IFPEN and University Grenoble Alpes in April 2016 in the OQUAIDO project (Optimisation et QUAntification d'Incertitudes pour les Données Onéreuses).

Stochastic spectral embedding for Bayesian inverse problems

P.-R. WAGNER
ETH Zurich

Supervisor(s): Prof. Bruno Sudret, Dr. Stefano Marelli

Ph.D. expected duration: Feb. 2017 - Feb. 2021

Adress: Stefano-Franscini-Platz 5 8093 Zürich

Email: wagner@ibk.baug.ethz.ch

Abstract:

The calibration of computational models based on collected measurements is a well-known problem in engineering and the applied sciences. A popular and powerful way to carry out this calibration is provided by the Bayesian inference framework. It reframes the calibration problem in the setting of updating *prior* information about the unknown model input parameters $\mathbf{X} \sim \pi(\mathbf{x})$ following the observation of measurements $\mathbf{y}$ through the well known Bayes' theorem:

$$\pi(\mathbf{x}|\mathbf{y}) = \frac{\mathcal{L}(\mathbf{x};\mathbf{y})\pi(\mathbf{x})}{Z}, \quad \text{with} \quad Z = \int_{\mathbb{R}^M} \mathcal{L}(\mathbf{x};\mathbf{y})\pi(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{1}$$

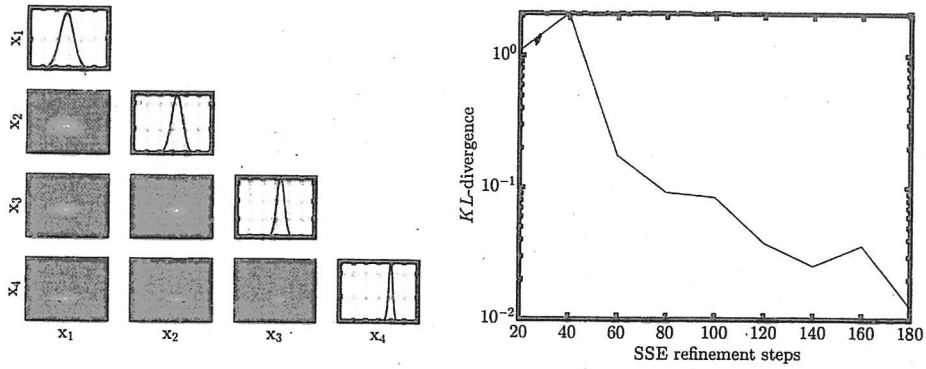
In this equation, the so-called posterior distribution $\pi(\mathbf{x}|\mathbf{y})$ reflects the updated information about the parameters $\mathbf{x}$. The connection between the prior and posterior distributions is established through the likelihood function $\mathcal{L}(\mathbf{x};\mathbf{y}) = \pi(\mathbf{y}|\mathbf{x})$.

The theoretical framework for updating the prior distribution is well established. In most real-world applications, however, it is not possible to analytically compute the posterior distribution. Nonetheless, numerous techniques to solve the Bayesian problem have been developed in the past. They range from classical methods that generate a sample distributed according to the posterior distribution (*Markov chain Monte Carlo* algorithms), to new methods that aim at finding a transport map to *push-forward* the uncertainty from the prior to the posterior distribution (*optimal maps* [2]) or likelihood-free methods (*approximate Bayesian computation* [1]).

It was shown in a recent publication [3] that the Bayesian updating problem can also be solved by constructing a polynomial chaos expansion (PCE) of the likelihood function $\mathcal{L}(\mathbf{x};\mathbf{y})$. The obtained spectral likelihood expansion (SLE) leads to analytical expressions for the normalization constant Z in Eq. (1), posterior moments and general quantities of interest. While these theoretical results are very promising, the method is not feasible in real-applications, as the required polynomial degree for a usable likelihood approximations typically exceeds the computational budget.

To reduce the required polynomial degree, we herein present a new algorithm titled *stochastic spectral embedding* that sequentially constructs low order PCEs of the likelihood function in increasingly smaller domains. The algorithm initially builds a PCE of the likelihood function over the whole prior support (identical to SLE). In further steps, the experimental design is enriched in subregions with high approximation errors and the unexplained residual is used to construct low order PCEs in these subregions. Due to the employed selective refinement, the procedure remains usable in moderate to high dimensions. An estimate of the generalization error is used to terminate the algorithm, once a sufficient approximation accuracy has been reached.

Due to the local orthogonality of the used polynomial basis functions, the normalization constant, marginals and moments of the posterior distribution can be computed analytically by post-processing the PCE coefficients. Furthermore, the ability of the SLE algorithm to analytically compute expectations of general functions w.r.t. the posterior distribution is conserved.



(a) Marginals $\pi(\mathbf{x}|\mathbf{y})$ after 180 refinement steps (b) Convergence of KL-divergence

Figure 1: Results of the presented SSE algorithm

To showcase the algorithm, a four dimensional inference problem is solved, where both the prior $\pi(\mathbf{x})$ and the likelihood function $\mathcal{L}(\mathbf{x};\mathbf{y})$ are given by multivariate normal distributions:

$$\pi(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu_{\text{prior}}, \Sigma_{\text{prior}}) \qquad \mathcal{L}(\mathbf{x};\mathbf{y}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{y}^{(i)}|\mathbf{x}, \Sigma). \tag{2}$$

The prior parameters μ_{prior}, Σ_{prior} and the measurement noise Σ of the likelihood function are known. Additionally, a set of $N = 10$ synthetic data measurements $\mathbf{y}^{(i)}$ is generated to complete the inference problem. The posterior distribution in this case is also given by a multivariate normal distribution and can be calculated analytically.

The SSE algorithm is applied to this problem using low degree polynomials of maximum degree $p = 3$. To judge convergence, the KL-divergence between the analytical posterior distribution and the approximation is computed after every 20 refinement steps. The results of the analysis are shown in Figure 1. The total number of likelihood evaluations in this study was $7,000$. The proposed algorithm converges well in the presented inversion problem. It is promising and will be tested in the future on more realistic cases.

References

[1] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[2] T. A. El Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[3] J. Nagel and B. Sudret. Spectral likelihood expansions for Bayesian inference. *Journal of Computational Physics*, 309:267–294, 2016.

Short biography – Paul-Remo Wagner is a third year PhD student with a Master's degree in civil engineering. His research focuses on Bayesian inversion and model calibration for engineering problems. He uses and develops advanced surrogate modeling tools to increase the practicality of Bayesian methods in real-world applications.

Conditional Quantile Optimization via Branch-and-Bound Strategies

LÉONARD TOROSSIAN
INRA (MIAT) - University of Toulouse (Institute of Mathematics)

Supervisor(s): Dr. Robert Faivre (INRA Toulouse), Prof. Aurélien Garivier (ENS-Lyon) and Dr. Victor Picheny (Prowler.io, Cambridge, UK)

Ph.D. expected duration: Nov. 2016 - Oct. 2019

Adress: INRA, 24 chemin de Borde Rouge, 31320 Auzeville-Tolosane

Email: leonard.torossian@inra.fr

Abstract:

We propose two branch-and-bound algorithms to optimize the conditional quantiles of *stochastic black boxes*. We consider systems that can be modeled as: $f : \mathcal{X} \times \Omega \to \mathbb{R}$, with f a reward or cost function, $\mathcal{X} \subset \mathbb{R}^D$ a design space and Ω a stochastic space. Contrary to *deterministic black boxes*, at a fixed x, the output is a random variable $Y_x = f(x, \omega)$ that follows an unknown distribution $\mathbb{P}(Y|X = x)$. We consider the classical setting of computer experiments: the function is only accessible through pointwise evaluations $f(x, \omega)$ and the gradient of any functional of f is unknown. Additionally, we assume that the variance of $\mathbb{P}(Y|X = x)$ may vary with respect to x (heteroscedasticity) and that $\mathbb{P}(Y|X = x)$ does not belong to any specific parametric class.

The conditional quantile function of order τ is defined as $q_\tau(x) = \inf\{q : F(q|X = x) \geq \tau\}$, $\tau \in (0, 1)$, where $F(\cdot|X = x)$ is the cumulative distribution function of $\mathbb{P}(Y|X = x)$, and we denote its supremum (that is assumed to be a maximum) as $q_\tau(x^*)$. Given a finite evaluation budget n, the goal of the presented algorithms is to propose a value $x(n)$ such that the *simple regret*

$$r_n = q_\tau(x^*) - q_\tau(x(n))$$

is as small as possible. To do so, the algorithms use a sequential strategy $x_1, \ldots, x_n$ that balances between exploration and intensification. Once the budget is over, the value with the best theoretical guarantees $x(n)$ is returned.

Classical bandit-based approaches [2] rely on an upper confidence bound (UCB) of the objective function instead of a simple estimator, that is, a surrogate function larger than the objective supremum with high probability. Computing an UCB everywhere on a continuous input space may be difficult. To facilitate the computation and guide the sampling strategy, the algorithms that we consider are based on recursive partitionings of $\mathcal{X}$ represented by hierarchical partition trees $\mathcal{T}$. The association between UCB and hierarchical trees has been widely used to optimize the conditional expectation [2, 5]. However, to our knowledge, we are the first to rely on this type of strategies in order to optimize conditional quantiles over continuous spaces.

We propose here two algorithms: Quantile Optimistic Optimization (QOO), which is an adaptation of the Deterministic Optimistic Optimization (DOO) [4], and Quantile Hierarchical Optimization (QHO), which is inspired from the Hierarchical Optimistic Optimization (HOO) [2]. Following the work of [6] on discrete problems, both rely on a deviation inequality for the empirical quantile [3] to build an UCB. The UCB function is designed to favor the leaves most likely to contain x^* in order to create an accurate estimator near the possible optimal points, while also favoring the leaves that have been least explored.

The principle of QOO is as follow. Starting from an initial partitioning $\mathcal{T}_1$ of $\mathcal{X}$, at each step t ($1 \leq t \leq n$) QOO computes an UCB for all the leaves. Then the leaf with the highest UCB is

selected and f is evaluated for an x chosen inside that leaf. If the quantile estimate associated to the leaf is accurate enough then the leaf is expanded into K children. Determining if the estimator associated to the leaf is accurate enough is a key point of QOO: this is achieved by associated a lower confidence bound (rely again on the deviation inequality [3]) to the UCB. Finally, $x(n)$ is chosen in the deepest node among those that have been expanded.

QHO differs both in the UCB computation and the tree growing process. Contrary to QOO, at each time t, QHO finds an "optimistic" path from the root to the leaf. That process implies the computation of an UCB at each depth h, based on all the subsequent children. Once a leaf is reached, QHO expands it immediately and f is evaluated in the K new leaves. At the end, $x(n)$ is chosen uniformly at random among the points that have been evaluated.

DOO only samples at the center of each leaf while HOO samples randomly inside the leaf, which leads in practice to different observation sets (with our without repeated observations). In our framework, both QOO and QHO can work with either sampling strategy.

Using analysis tools from the bandit literature [5, 1], we prove non asymptotic guarantees for our algorithms, assuming only a smoothness relative to the optimal point:

$$q_\tau(x^*) - q_\tau(x) \leq A \left\|x^* - x\right\|^\beta, \quad \beta > 0, \quad A > 0.$$

We finally provide an empirical analysis that illustrates the respective merits of our algorithms. In particular, we discuss the benefits and drawbacks of the use of repetitions in the sampling strategies.

References

[1] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[2] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

[3] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[4] Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in neural information processing systems*, pages 783–791, 2011.

[5] Rémi Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

[6] Balazs Szorenyi, Róbert Busa-Fekete, Paul Weng, and Eyke Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *32nd International Conference on Machine Learning*, pages 1660–1668, 2015.

Short biography – I graduated from Université Pierre et Marie Curie in 2016 with a MSc in Modelling and Optimization. I started my PhD at INRA and IMT in November 2016 co-directed by R. Faivre, V. Picheny and A. Garivier. My work is funded by MIA and the Occitanie region. My PhD subject is about metamodeling and robust optimization of stochastic black boxes, I aim to build a tool able to take optimal decisions under risk aversion.

MASCOT-NUM 2019

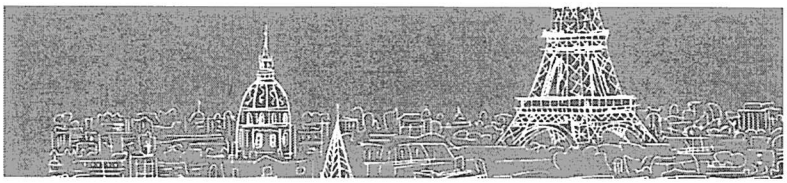
COSME Emmanuel
Université Grenoble Alpes

Abstract

A short introduction to data assimilation (Course 1/2)

With this contribution I propose a short introduction to data assimilation (DA), with a strong emphasis on the methods used in geosciences (my expertise). The first part will be dedicated to a global and intuitive perspective: which problems are solved using DA, and how. In the second part, I will expose the maths. The main methods reviewed are: the particle filter, the Kalman filter, the EnKF, and 4DVar, which are mostly used in geophysics. The third and last part will present examples and typical future challenges in geophysical DA.

MASCOT-NUM 2019



MÜLLER Werner
Johannes Kepler University Linz

Abstract

Privacy sets revisited

In computer simulation experiments, which have now become a popular substitute for real experiments, one usually aims to spread out the measurements uniformly across the design space, yielding so-called space-filling designs. Most of the literature on space-filling designs attempts to achieve its aim by optimizing a prescribed objective measuring a degree of space-fillingness (see eg. Pronzato and Müller, 2012). These criteria are sometimes combined with an estimation or prediction oriented criterion. Let us label those as "soft" space-filling methods. In contrast "hard" space-filling methods ensure desirable properties by enforcing constraints on the designs, as for instance provided by privacy sets (see Benková et al. 2016), such that a secondary criterion can be used for optimization. External constraints such as on the design region or else can be incorporated in a similar manner.

This talk provides a fresh look on the role of privacy sets for the construction of space-filling designs with new algorithms and new examples. In contrast to the privacy sets considered in our previous work, the new constraints guarantee some minimal distance between any two design points, which spreads out the measurements across the design space in a very natural way.

References:

Eva Benková, Radoslav Harman, and Werner G. Müller. Privacy sets for constrained space-filling. Journal of Statistical Planning and Inference, 171:1-9, 2016.

Luc Pronzato and Werner G. Müller. Design of computer experiments: space filling and beyond. Statistics and Computing, 22(3):681-701, 2012.

MASCOT-NUM 2019

METELKINA Asya
Azoth Systems

Abstract

Finding a compromise between information and regret in clinical trials

Joint work with Luc Pronzato. We consider the treatment allocation problem in the setting of comparative clinical trials in which patients arrive sequentially. For this type of trials, response adaptive and/or covariate-adjusted designs can be used to construct the sequence of allocation probabilities to satisfy the study objectives. The most common target for allocation designs is a maximum precision of the estimation of treatment's model parameters, e.g. in some generalized linear models. This can be achieved through the maximization of an information criterion, a concave function of the Fisher information matrix. But ethical clinical study should also reduce the number of patients who receive inferior treatments. We propose a compromise criterion defining a trade-off between information and ethics objectives through the convex combination of an information criterion and a regret function. Under mild conditions, we show the existence and give an explicit construction of the locally optimal (maximizing the compromise criterion) allocation measure on the space of covariates. This allocation measure is then used in an oracle covariate-adaptive allocation procedure. However the construction of the optimal allocation measure is complicated and requires an a prioriknowledge of covariates distribution. We show how these difficulties can be avoided by using a covariate-adaptive allocation rule based on empirical allocation measures, which we show to converge to the optimal measure. To deal with the unknown model parameters, we propose a response-adaptive allocation rule that uses current Maximum Likelihood estimates of model parameters. Comparison of our allocation designs with recently proposed adaptive designs from literature will be given.

MASCOT-NUM 2019

ZHIGLIAVSKY Anatoli
Cardiff University

Abstract

Measures minimizing regularized dispersion

We consider a continuous extension of a regularized version of the minimax, or dispersion, criterion widely used in space-filling design for computer experiments and quasi-Monte Carlo methods. We show that the criterion is convex for a certain range of the regularization parameter (depending on space dimension) and give a necessary and sufficient condition characterizing the optimal distribution of design points. Using results from potential theory, we investigate properties of optimal measures. The example of design in the unit ball is considered in details and some analytic results are presented. Using recent results and algorithms from experimental design theory, we show how to construct optimal measures numerically. They are often close to the uniform measure but do not coincide with it. The results suggest that designs minimizing the regularized dispersion for suitable values of the regularization parameter should have good space-filling properties. An algorithm is proposed for the construction of n-point designs.

MASCOT-NUM 2019

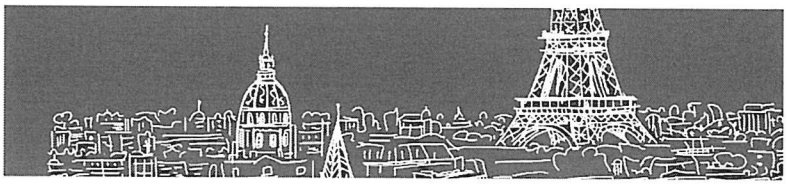
VAZQUEZ Emmanuel
CentraleSupelec

Abstract

Bayesian sequential strategies for computer experiments

In this talk, I will briefly present our past research work on Bayesian sequential strategies for computer experiments. Stepwise Uncertainty Reduction and Bayesian optimization will be the main topics that I will address here.

MASCOT-NUM 2019



BATES Ron
Rolls Royce

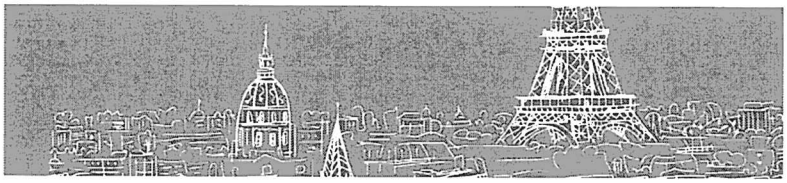
Abstract

Model-Based Robust Design in Industry

The need for Robust Design has led to many innovations in computer experiments which itself has driven advances in automation and parametrization of simulation models. Now, as empirical data become cheaper to acquire and more ubiquitous, the challenge is to combine both physics-based and data-driven approaches to support the product development process.

This talk will attempt to give an end-to-end account of how Robust Design supports the decision-making process from component development through to the validation of system-level requirements.

MASCOT-NUM 2019



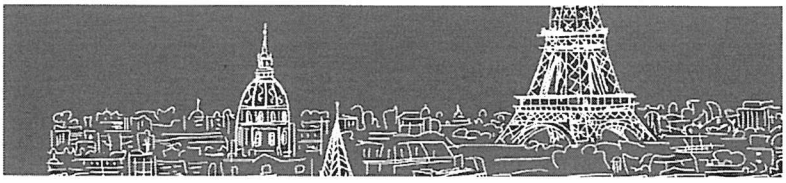
WOLSZTYNSKI Eric
University College Cork

Abstract

Tumour characterization from Positron Emission Tomography imaging data

This work focuses on the development of statistical methods for the analysis of cancer imaging data. We consider in particular problems related to the assessment of prognosis, staging and disease recurrence, mainly from Positron Emission Tomography (PET) imaging data but also from MRI and CT modalities. In particular, spatial heterogeneity of the 18F-fluorodeoxyglucose uptake pattern in PET has been established as a strong prognostic indicator for sarcoma, lung, breast and other cancers. Our approach consists in developing new quantification methodologies for characterization of tumour metabolism and structure. Spatial models of the volumetric distribution of PET tracer uptake within the volume of interest are used to extract relevant metabolic and structural descriptors of the tumour. These variables are then considered for the assessment of prognosis and therapeutic response. This work involves a number of technical aspects from various areas including nonparametric estimation, regularization and statistical learning. The main application of this research is cancer patient-adaptive treatment, but it also links with problems found in other biomedical and actuarial applications.

MASCOT-NUM 2019



JOURDAN Astrid
EISTI

Abstract

Space-filling designs based on Rényi entropy

Space-filling designs are commonly used for selecting the input values of time-consuming computer codes. Since the true relation between the computer response and inputs is not known, the designs should allow one to fit a variety of models and should provide information about all portions of the experimental region. One strategy for selecting the values of the inputs which to observe the response is to choose these values so they are spread evenly throughout the experimental region, according to a "space-filling criterion". Many space-filling criteria have been investigated in the literature. Some of them quantify how the points fill up the space using the distance between points, such as the maximin distance [5] or the Audze-Eglais criterion [1]. Others measure the difference between the empirical distribution of the design points and the uniform distribution, such as the discrepancy ([9], [3]) or Kullback-Leibler criterion [6].In this paper, we use results discussed in Pronzato's work ([10], [8]) to build space-filling designs based on Renyi's entropy. Suppose that the points $x_1,...,x_n$ of the design D, are *n* independent observations of the random vector $X=(X_1,...,X_d)$ with absolutely continuous density function *f* concentrated on the unit cube $[0,1]^d$ (we reduce the design space to the unit cube). Rényi entropy,

$$H_q(D) = \frac{1}{1-q} \ln \int_E f(x)^q dx \quad \text{, with } q \in]0,1[$$

measures the difference between f and the uniform density function in so far as, one always has Hq(D) 0 and the maximum value of Hd(D), zero, being uniquely attained by the uniform density. This latter property confirms that maximizing Rényi entropy makes f converge toward the uniform density. We investigate three ways for estimating the entropy:

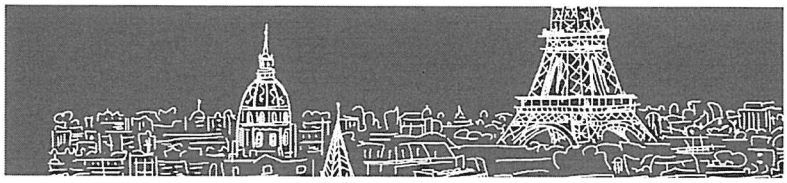
- a Monte Carlo method [2] where the unknown density function f is replaced by its kernel density estimate [11],• an estimation based on the nearest neighbor distance [7],• a method based on the minimum spanning tree built from the design points [4].

References

[1] Audze, P. and V. Eglais. New approach for planning out of experiments, Problems of Dynamics and Strengths, 35:104-107, 1977.[2] Beirlant J., Dudewicz E.J., Györfi L., Van Der Meulen E.C. Nonparametric entropy estimation : an overview. Int. J. Math. Stat. Sci., 6(1):17-39, 1997.[3] Fang K.T., Li R., Sudjianto A. Design and modeling for computer experiments. Chapman&Hall, London, 2006.[4] Hero A., Bing Ma, Michel O., Gorman J. Applications of entropic spanning graphs. Signal Processing Magazine, IEEE, 19:85 – 95, 2002.

[5] Johnson M.E., Moore L.M., Ylvisaker D. Minimax and maximin distance design. J. Statist. Plann. Inf., 26:131-148, 1990.[6] Jourdan A. et Franco J. Optimal Latin hypercube designs for the Kullback-Leibler criterion. AStA Advances in Statistical Analysis, 94(4):341-351, 2010.[7] Kosachenko L.F., Leonenko N.N.. Sample estimate of entropy of a random vector. Problem of Information Transmission, 23:95-101, 1987.[8]Leonenko N., Pronzato L.,Savani V., A class of Rényi information estimators for multidimensional densities, Ann. Statist. 36:2153 - 2182; correction by Leonenko and Pronzato 2010, Ann. Statist., 38:3837 – 3838, 2008.[9]Niederreiter H. Point sets and sequences with small discrepancy. Monasth. Math., 104:273-337, 1987.[10] Pronzato L. Minimax and maximin space-filling designs: some properties and methods for construction. Journal de la Société Française de Statistique, 158(1), 2017.[11]Silverman B.W. Density estimation for statistics and data analysis. Chapman & Hall, London, 1986.

MASCOT-NUM 2019



ELSHEIKH Ahmed
Heriot-Watt University

Abstract

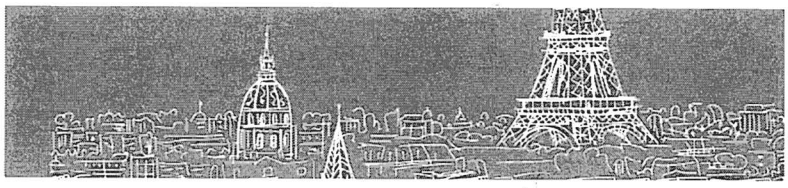
Synthesis of geological images using deep learning techniques

We propose a framework for synthesis of geological images based on an exemplar image (a.k.a. training image). We synthesize new realizations such that the discrepancy in the patch distribution between the realizations and the exemplar image is minimized. Such discrepancy is quantified using a kernel method for two-sample test called maximum mean discrepancy. To enable fast synthesis, we train a generative neural network in an offline phase to sample realizations efficiently during deployment, while also providing a parametrization of the synthesis process. We assess the framework on a classical benchmark of a binary image representing channelized subsurface reservoirs. Results show that the method is effective in reproducing the visual patterns and spatial statistics (image histogram and two-point probability functions) of the exemplar image, providing a promising direction towards parametric synthesis of geology directly from an exemplar image.

References:

[1] Shing Chan, Ahmed H. Elsheikh, "Exemplar-based synthesis of geology using kernel discrepancies and generative neural networks", preprint arXiv:1809.07748. URL: https://arxiv.org/abs/1809.07748

MASCOT-NUM 2019



CELSE Benoit
IFPEN

Abstract

Machine Learning techniques in Industry. Application to oil refining

In recent years, the increasing integration of the Internet of Things into production industry is at the genesis of a new digital industrial revolution known as Industry 4.0 [1]. The core component of Industry 4.0 is the concept of the digital twin. The main objectives of digital twin are to replace or reduce expensive, time-consuming physical experiments with rapid, inexpensive computer simulation [2]. Accurate and reliable predictive models for physiochemical properties of Hydrocracking process (HCK) products are extremely important. It can help petroleum refinery industries to save time and expansion on costly experiments. In our case, this is the main motivation to synthesize the available knowledge base of HCK process to build a digital twin capable of predicting the product properties of valuable petroleum fractions based on scientific principles. However, for the efficient execution of digital twins is it required to use the different steps of the Knowledge Discovery in Databases-process (KDD-process) [3]. That means automatic extraction of non-obvious, hidden knowledge from large volumes of data [4]. Ideally, the twin would enable:

1. Data cleaning and preprocessing to handle:
   - missing data items by deletion or imputation approach,
   - unexpected values (variables under consideration are expected to have values within a predefined range) by expert pre-processing.
2. Outliers Detection to identify and remove unwanted samples from data. In this work, we use the Local Outlier Factor (LOF) technique [5].
3. Variables selection to select optimum number of variables from a large pool of variables. In the present work we applied leaps [6] and Random Forests [7] algorithms to determine suitable descriptors.
4. Machine learning to build models that characterize the impact of different physicochemical properties of the product. For this, Linear Regression, Kriging ,Support Vector Regression , Random Forest and Gradient Boosting Machine are proposed

5. Validation of the models using either a dedicated database or cross-validation techniques.

The proposed twin was tested to predict the specific gravity of the vacuum gasoil (VGO), diesel and heavy naphtha cuts in Mild HCK (process that uses low to intermediary pressures and relatively low conversions). The results from this work will be presented. Results are promising but many unforeseen difficulties had to be addressed. In the future, the proposed framework can also be used to predict other properties (cetane number, sulphur and nitrogen content, etc.) with several process (High Pressure Hydrocracking, FCC ...). Furthermore, this type of methodology is also extended to predict industrial deactivation.

References

[1] F. Shrouf, J. Ordieres, G. Miragliotta. Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm. 2014 IEEE International Conference on Industrial Engineering and Engineering Management, 2014, 697-701.

[2] Grieves M., Vickers J. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems, in Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches. Éd. F.-J. Kahlen, S. Flumerfelt, A. Alves. Springer International Publishing, Cham, 2017, 85-113.

[3] Schuh G., Rudolf S., Riesener M., others. Design for Industrie 4.0. 2016 14th International Design Conference, 1387-1396.

[4] Fayyad U., Piatetsky-Shapiro G., Smyth P. The KDD Process for Extracting Useful Knowledge from Volumes of Data, Commun. ACM, 1996, 39, 11, 27-34. DOI: 10.1145/240455.240464.

[5] Ding H., Ding K., Zhang J., Wang Y., Gao L., Li Y., Chen F., Shao Z., Lai W. Local outlier factor-based fault detection and evaluation of photovoltaic system, Solar Energy, 2018, 164, 139-148. DOI: 10.1016/j.solener.2018.01.049.

[6] Feorge M. Furnival, Robert W. Wilson. Regressions by Leaps and Bounds, 16, 1974.

[7] Genuer R., Poggi J.-M., Tuleau-Malot C., Villa-Vialaneix N. Random Forests for Big Data, Big Data Research, 2017, 9, 28-46. DOI: 10.1016/j.bdr.2017.07.003.

MASCOT-NUM 2019

MAZO Gildas
INRA

Abstract

An optimal balance between explorations and repetitions in sensitivity analysis

Sensitivity analysis is well developed for deterministic computer models. When the computer model is stochastic, however, it is less clear what its performance is, or even what it means. Typically, the computer model is repeated several times, say m, at each one of the explorations of the input space, leading to a number of runs proportional to mn, where n is the number of explorations. How does the performance of the estimators depend on (n,m)? What is the definition of a Sobol index for stochastic models? Our contribution is twofold. First, we build a formalism in which two definitions of Sobol's indices are given, estimators are built and asymptotic properties are established in terms of both n and m, revealing the differences between the two kinds of sensitivity indices. Second, we address the problem of choosing an optimal balance between repetitions and explorations under a fixed budget constraint. We define the optimal number of repetition as the couple (n,m) that minimizes the risk of a bad ranking of the input factors. We proceed by designing a two-step procedure in which the first step serves to estimate the optimal number of repetitions and the second step the sensitivity indices based on the number found in the first step. We show that this procedure is asymptotically oracle, in the sense that it does arbitrarily as good as the procedure in which the optimal number of repetitions is known. Numerical illustrations are provided to illustrate our theoretical findings.

MASCOT-NUM 2019

MOREAUD Maxime IFPEN

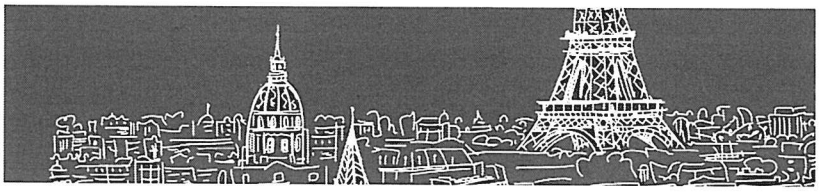
Abstract

Efficient Topological and morphological characterization of 3D complex microstructures

Abstract: Porous media characterization is central for heterogeneous catalysis for the production of biofuels and chemical intermediates by biomass transformation. Their description should provide certain connection to some of their physicochemical properties, and concerning their activity or selectivity.Standard geometric descriptions, such as porous volume fraction, granulometry of pores, or specific surface area, are seldom sufficient for this purpose. This is why we have developed new morphological and topological descriptors using the so-called "distance transform" with adapted time-efficient numerical methods. The present work is a global attempt to provide a realistic description of the microstructure of porous media; it should help to define an optimal microstructure modelization taking into account intended textural and usage properties. Such a description can also lead to a structural classification of porous media. We will present a first approach addressing the ability for given particle's sizes to go through the porous network until a critical radius. Then, we will define a new versatile tortuosity descriptor based on the travel distance of a particle in a porous maze. The computation of these new descriptors will be shown using plug im!, a signal and image processing open access software, on several types of porous media such as zeolites, metal-organic frameworks and alumina catalyst supports.

Maxime Moreaud*†, Johan Chaniot*, Thierry Fournel', Jean Marie Becker', Loïc Sorbier**IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize, France'Université de Lyon, Université Jean Monnet de Saint Etienne, CNRS UMR 5516, Laboratoire Hubert Curien, F-42000 Saint Etienne, France† MINES ParisTech, PSL-Research University, CMM, 35 rue Saint Honoré, 77305 Fontainebleau, Francemaxime.moreaud@ifpen.fr

MASCOT-NUM 2019



LE RICHE Rodolphe
CNRS LIMOS at Mines Saint-Etienne

Abstract

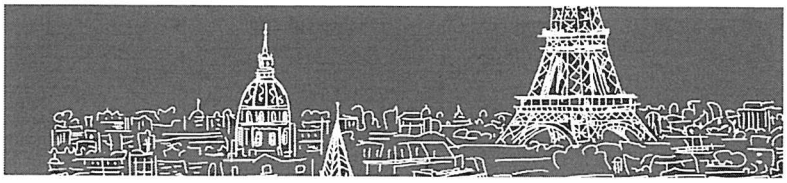
Multi-objective Bayesian optimization with reference point

Bayesian algorithms (e.g., EGO, GPareto) are a popular approach to the mono and multi-objective optimization of costly functions. Despite the gains provided by the Gaussian models, convergence to the problem solutions remains out of reach when the number of variables and / or the number of objective functions increase.
In this presentation, we show how with Gaussian processes it is possible to restrict ambitions in order to recover problems that can be solved.

With strong restrictions on the number of objective function evaluations, it is often only feasible to target a specific point of the Pareto front. We describe the mEI criterion to do so. When no such point is known a priori, we propose to target the Pareto front center. Thus, we define this center, explain how to estimate it and how to detect convergence to it.
Once the center of the Pareto front has been found, we propose to enlarge the search for Pareto optimal solutions around it in a manner that is compatible with the remaining computational budget. To achieve this, virtual Bayesian optimizations are carried out on the Gaussian processes.
Finally, we discuss how to parallelize the resulting multi-objective Bayesian optimization algorithm.

This talk summarizes a joined work with David Gaudrie and Victor Picheny.

MASCOT-NUM 2019



LECOMTE Jean-François
IFPEN

Abstract

Statistical 3D analysis of foam bubbles in poreous media using a large NoSQL Database

Modeling foam behavior in porous media is a complex task. X-ray computed tomography experiments are used to obtain at least an accurate visual description of physical processes. Thus, it provides a large collection of 3D images. Using python and scikit-image, an automatic sequence of operation able to process these images was developed to isolate foam bubbles in porous media and to extract geometric and statistical features over this huge set of images. To transform such a large collection of raw images into meaningful features requires also a clean upstream preparation. Therefore, a database was also designed in order to let python apply standard algorithms of image processing in an automated way.

MASCOT-NUM 2019

ZAHM Olivier
INRIA

Abstract

Detecting and exploiting the low-effective dimension of multivariate problems using gradient information

Approximation of multivariate functions is a difficult task when the number of input variables is large. Identifying the directions where the function does not vary significantly is a key preprocessing step to reduce the complexity of the approximation algorithms. We propose a gradient-based method that permits to detect such a low-dimensional structure of a function. The methodology consists in minimizing an upper-bound of the approximation error obtained using Poincaré-type inequalities. This generalizes the active subspace method to vector-valued functions. We also show the connection with standard screening techniques used in Global Sensitivity Analysis. Finally, the method naturally extends to non-linear dimension reduction, e.g. when the function is not only constant along a subspace but along a low-dimensional manifold.

MASCOT-NUM 2019

GUITON Martin
IFPEN

Abstract

Optimization and reliability design of a floating offshore wind turbine

The floating wind turbine technologies currently under development must be designed to withstand environmental conditions for several decades, taking into account several uncertainties on the solicitations and the models. According to design standards, the validation of a configuration must satisfy in particular both extreme limit state and fatigue limit state. A reduction of the cost of electricity generated by these turbines is required to become comparable with other sources of power generation, thus motivating optimization of the configuration.
Both the evaluation of the reliability constraints, and a fortiori the optimization submitted to these constraints, constitute a challenge because of the considerable computation cost. This cost results from the complexity of the aero-hydro-servo-dynamic simulators as well as from the very large number of load cases prescribed by the design standards. After introducing the problem, we present several strategies to limit this cost calculation. The calculation of the constraints in extreme limit state can be simplified by describing the input signal of the loading (swell, wind) by harmonics with a hundred random variables. Assuming the load process to be stationary, we recover a time independent reliability problem. The computation of the most probable point at an arbitrary time, greater than initial transient stage, enables to determine the critical loading with reduced simulation times when compared to standards. The outcrossing rate can be calculated with a limited cost in a FORM framework or more precisely with dimension reduction strategies. Results are illustrated for the case of a mast of a wind turbine. The calculation of fatigue stresses can be considerably accelerated by constructing a response surface based on an optimal experimental design. In the case of optimization, we illustrate the interest of a non-derivative algorithm (SQA), developed at IFPEN, which is particularly adapted to this type of simulator, with the application to the configuration of an electrical cable connecting floating wind turbines. Finally, we propose lines of thought to decouple the optimization loop from the calculation configuration of the reliability constraints. This last point is addressed in a thesis to optimize the configuration of the mooring lines for a floating wind turbine.

M. Guiton, T. Perdrizet, N. Delépine, Y. Poirette, G. Huwart IFPEN Direction Mécanique Appliquée M. Munoz-Zuniga, A. Cousin, D. Sinoquet IFPEN Direction Mathématique Appliquée J. Garnier, CMAP Polytechnique

Poster sessions

Surrogate modeling of stochastic simulators using Karhunen-Loève expansions

S. Azzi
Télécom ParisTech, LTCI, Université Paris-Saclay

Supervisor(s): J. Wiart (Télécom ParisTech), B. Sudret (ETH Zürich)

Ph.D. expected duration: Oct. 2016 - Oct. 2019

Adress: Télécom ParisTech, 46 Rue Barrault 75013 Paris

Email: soumaya.azzi@telecom-paristech.fr

Abstract: In engineering problems, simulators commonly contain sources of uncertainty due to measurements for example. They are called stochastic simulators because they yield a probability density function (PDF) with respect to every input. Even though through numerical computations, stochastic simulators can be investigated, they remain computationally expensive. Metamodels are mathematical functions that mimic the behavior of simulators and are used to overcome the pricey calls to the simulators. The abstract introduces a metamodeling approach for stochastic simulators based on Karhunen-Loève (KL) expansion [4].

Let $H(x,\omega) \in \mathbb{R}$ be a stochastic process on $D \times \Omega$, where $x \in D \subset \mathbb{R}^n$ and ω in the sample space Ω. The stochastic simulator is modeled as a stochastic simulator, and its surrogate is a stochastic process as well, noted $\hat{H}(x,\omega)$. Let the stochastic process $H(x,\omega)$ be a zero mean second order process. Its covariance operator is denoted $C(x,y)$ and let λ_i and ϕ_i be respectively its eigenvalues and eigenvectors. Then the KL expansion [4] reads as follow :

$$H(x,\omega) = \lim_{p \to \text{inf}} \sum_{i=1}^{p} \sqrt{\lambda_i}\xi_i(\omega)\phi_i(x) \tag{1}$$

In practice, several calls to the stochastic simulator are made, let M be the size of the design of experiment set (DoE) and M the number of realization on each point from the DoE. The simulated process is then a matrix with M rows and N columns, each row represents the N realizations made over a point from the DoE. Each column represents a trajectory, meaning that for all $x \in DoE$, simulations were carried with a same seed. Based on the data from the simulation, the empirical covariance matrix is evaluated, ϕ_i and λ_i are calculated. The aim is to predict the PDF of a new point $x^* \in D$. Based on Eq.1, the predicted response reads as $\hat{H}(x^*,\omega) = \sum_{i=1}^{M} \sqrt{\lambda_i}\hat{\xi}_i(\omega)\hat{\phi}_i(x)$.

- $\phi_i(x^*)$ is unknown, the eigenvectors are only computed for the DoE set. To overcome this limitation, a ϕ_i originally known only over the M point of the DoE can be interpolated to predict $\phi_i(x^*)$.
- Alternatively, a metamodel of the covariance is build using polynomial chaos expansions [1], the eigendecomposition is then performed to get $\hat{\phi}_i$. Both ways are used and compared.
- Concerning the random variables, they are the projection of H onto the base of the eigenvectors $\hat{\phi}_i$, $i \in \{1 \dots M\}$ [4]

$$\hat{\xi}_i(\omega_k) = \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^{M} H(x^{(j)},\omega_k)\hat{\phi}_i(x^{(j)}) \tag{2}$$

This approach was applied to a radio frequency exposure (RF) computational simulator, the simulator evaluates the exposure to RF waves of a population living in cities [3]. The inputs of this simulator are the parameters of the city, and the output is the RF exposure of the population. The simulator is stochastic, given a set of city parameters, numerous cities can be generated [2] (Figure 1) hence numerous exposure rates can be evaluated (for the same city parameters).

Figure 1: Three realizations of virtual cities showing the same parameters (street width, anisotropy, building height and length).

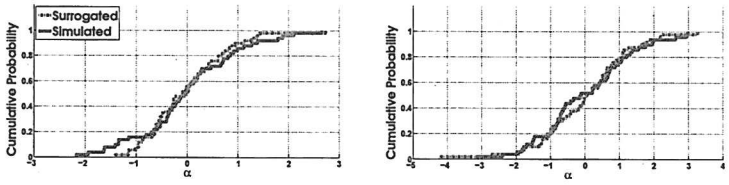


Figure 2: Surrogated and simulated CDF's plotted when the eigenvectors are linearly interpolated (right) and when the covariance is interpolated using PCE (left).

The stochastic city simulator was simulated on 30 points, 50 times. A cross validation was applied with 10% of the data left to the test (Figure 2). The accuracy of the model is evaluated by comparing the 'real' PDF and the surrogated one using different metrics : histogram intersection, Hellinger distance and Jensen Shannon divergence.

For the sensitivity analysis (SA) and based on the simulations, the entropy on each point of the DoE, noted $\mathcal{E}(x)$ is evaluated, a metamodel $\hat{\mathcal{E}}(x)$, $x \in D$ is build. The SA approaches can be evaluated on $\hat{\mathcal{E}}$.

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] T. Courtat, L. Decreusefond, and P. Martins. Stochastic simulation of urban environments. application to path-loss in wireless systems. *arXiv:1604.00688,*, 2016. [Online].

[3] Y. Huang and J. Wiart. Simplified assessment method for population RF exposure induced by a 4G network. *IEEE Journal of Electromagnetics, RF, and Microwaves in Medicine and Biology*, 1:34–40, 2017.

[4] Roger G. Ghanem Pol Spanos. *Stochastic Finite Elements: A Spectral Approach.*

Short biography – Azzi Soumaya is a Ph.D. student within Chaire C2M, LTCI, Télécom ParisTech. Her research interests include stochastic computation, surrogate modeling and uncertainty quantification. She received the M.Sc. degree in applied mathematics from Blaise Pascal University, Clermont Frd, France.

Interpretability of statistical learning models in an industrial context

CLEMENT BENARD
Safran Tech, Sorbonne Universite

Supervisor(s): S. Da Veiga (Safran Tech), E. Scornet (CMAP, Ecole Polytechnique), G. Biau (LPSM, Sorbonne Universite)

Ph.D. expected duration: Dec. 2018 - Nov. 2021

Adress: Safran Tech, 1 rue Genevieve Aube, 78114 Magny-les-Hameaux

Email: clement.benard@safrangroup.com

Abstract:

In the manufacturing industry, the core of production processes involves complex physical and chemical phenomena. Their control and efficiency is of critical importance. In practice, data is collected along the manufacturing line, characterizing both the production conditions and its quality. State-of-art supervised learning algorithms can successfully catch patterns of such complex physical phenomena, characterized by non-linear effects and low-order interactions between parameters. However, any decision impacting a production process have long term and heavy consequences, and then, cannot simply rely on stochastic modeling. A deep physical understanding is required and black-box models are not appropriate. Models have to be interpretable, i.e. provide an understanding of the internal mechanism that build a relation between inputs and ouputs, to provide insights to guide the physical analysis. There is no agreement in statistics and machine learning communities about a rigorous definition of interpretability [6]. It is yet possible to define minimum requirements for interpretability: simplicity, stability [8] and predictivity.

Decision tree [2] can model highly non-linear patterns while having a simple structure and is then widely used when interpretability is required. Decision tree is also highly unstable to small data pertubation, which is a very strong limitation to its practical use. Random forest [1] stabilizes decision trees by aggregating many of them, it strongly improves accuracy but the model is a black box. Another class of supervised learning method can model non-linear patterns while having a simple structure: rule models. A rule is a conjunction of constraints on inputs variables that form a hyper-rectangle in the input space, where the estimated output is constant. A collection of rule is combined to form a model. Many algorithms were developped, among them: SLIPPER [3], Rulefit [4], Node Harvest [7] and BRL [5]... They share the same draw brack as trees: instability.

In this work, we design a new classification algorithm which inherits the accuracy of random forests, the simplicity of decision trees while having a stable structure for problems with low-order interaction effects. The principle of random forest is used, but instead of aggregating predictions, we focus on the probability that a given hyper-rectangle (a node) is contained in a randomized tree. The nodes with the highest probabilities are robust to data pertubation and represent strong patterns. They are selected to form a stable rule ensemble model. Our proposed algorithm works as follows:

1. Bin data using empirical quantiles.
2. Generate a large number of rules with the random forest procedure.
3. Select rules based on their frequency of appearance in the random forest.
4. Average the selected rules to form a rule ensemble model.

Many simulations on public datasets of the UCI repository (Asuncion and Newman 2007) show good performance of the procedure in terms of both predictive accuracy and stability.
We use 1 - AUC to measure accuracy. To evaluate stability, a 10-fold cross-validation is run, and, for each pair of folds the relative size of the intersection between the two lists of rules is computed.

Dataset	Random Forest	Rulefit	Node Harvest	CART	Our Method
Diabetes	0.17	0.18	0.19	0.22	0.18
Heart Statlog	0.10	0.11	0.12	0.17	0.14
Heart C2	0.10	0.11	0.12	0.19	0.10
Heart H2	0.11	0.10	0.11	0.18	0.12
Credit German	0.21	0.23	0.26	0.30	0.24
Credit Approval	0.07	0.06	0.07	0.11	0.07
Ionosphere	0.03	0.06	0.07	0.12	0.12
Breast Wisconsin	0.01	0.02	0.02	0.05	0.01

Table 1: Accuracy

Dataset	Node Harvest	CART	Our Method
Diabetes	60%	28%	76%
Heart Statlog	50%	34%	65%
Heart C2	55%	28%	56%
Heart H2	45%	37%	72%
Credit German	47%	39%	73%
Credit Approval	30%	24%	59%
Ionosphere	33%	24%	77%
Breast Wisconsin	60%	57%	82%

Table 2: Stability

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. *Wadsworth International Group*, 1984.

[3] William W Cohen and Yoram Singer. A simple, fast, and effective rule learner. *AAAI/IAAI*, 99:335–342, 1999.

[4] Jerome H Friedman, Bogdan E Popescu, et al. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

[5] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[6] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[7] Nicolai Meinshausen. Node harvest. *The Annals of Applied Statistics*, pages 2049–2072, 2010.

[8] Bin Yu et al. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

Short biography – Clement Benard is a research engineer at Safran Tech and a first year PhD in statistics, in collaboration with LPSM, Sorbonne Universite.

Gaussian process metamodeling for functional-input coastal flooding code

J. BETANCOURT
Toulouse Mathematics Institute, University Paul Sabatier, France

Supervisor(s): Prof. Thierry Klein (Toulouse Mathematics Institute), Prof. François Bachoc (Toulouse Mathematics Institute)

Ph.D. expected duration: Aug. 2017 - Sep. 2020

Adress: Toulouse Mathematics Institute, 118 Route de Narbonne, 31400 Toulouse

Email: jbetanco@math.univ-toulouse.fr

Abstract:

Floods in general affect more people than any other hazard with 1.5 billion people affected in the last decade of the 20th century. Many recent events (e.g.: Katrina, USA 2005; Xynthia, France 2010) illustrate the complexity of coastal systems and the limits of traditional forecast and early warning systems and flood risk analysis. Recent scientific progresses now allow properly modeling coastal flooding events. Such models are nevertheless very expensive in terms of computation time (multiple hours) which prevents any use for forecast and warning. A widely used method to approach this type of limitation is to build a reduced model (often called surrogate model or metamodel), able to provide high precision estimates of the response surface at an acceptable computation time.

This study was developed in the frame of the ANR RISCOPE project, which studies the development of metamodels for coastal flooding early warning [1]. In the coastal flooding context, the metamodel should be able to deal with functional inputs associated to time varying maritime conditions such as the tide and surge. Among all the types of metamodels available (polynomials, splines, neural networks, etc.), we focus on Kriging (also called Gaussian process model), characterized by its mean and covariance functions [6, 5]. The main advantage of Kriging is its ability to provide both a prediction of the computer code and the uncertainty attached to this prediction [3]. Kriging metamodels were originally developed for scalar inputs, however, they can be also built for functional inputs. To this end, we project each functional input on a functional basis, and then we use the projection as inputs of the metamodel. In this study we compare results of metamodels based on B-splines [2] and PCA [4] projection methods, as well as two different forms to meassure the distance among functional basis within the covariance function: (i) taking the coefficients of the functional decomposition as independent scalar inputs; and (ii) using an adapted distance for functional decompositions. We illustrate a procedure for identification of relevant functional inputs for the metamodel. We further discuss two approaches to tune the dimension of the projections: i) based on the error of the projection; ii) based on the performance of the metamodel. Our results show that the approach based on the error of the projection, being the most widely used in the literature, may lead to unnecessarily large projection dimensions. In contrast, the approach based on metamodel performance presents the virtue of directly pointing to the final objective of building a fast and accurate metamodel. All codes were implemented in R and the metamodel was validated through a case study based on real data gathered at Gâvres coast in France.

References

[1] Anr riscope project. https://perso.math.univ-toulouse.fr/riscope/. Accessed: 2018-12-04.

[2] Thomas Muehlenstaedt, Jana Fruth, and Olivier Roustant. Computer experiments with functional inputs and scalar outputs by a norm-based approach. *Statistics and Computing*, 27(4):1083–1097, 2017.

[3] Simon Nanty, Céline Helbert, Amandine Marrel, Nadia Pérot, and Clémentine Prieur. Sampling, metamodeling, and sensitivity analysis of numerical simulators with functional stochastic inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):636–659, 2016.

[4] Simon Nanty, Céline Helbert, Amandine Marrel, Nadia Pérot, and Clémentine Prieur. Uncertainty quantification for functional dependent random variables. *Computational Statistics*, 32(2):559–583, 2017.

[5] Carl Edward Rasmussen and Christopher KI Williams. Gaussian processes for machine learning. 2006. *The MIT Press, Cambridge, MA, USA*, 38:715–719, 2006.

[6] Thomas J Santner, Brian J Williams, and William I Notz. *The design and analysis of computer experiments*. Springer Science & Business Media, 2013.

Short biography – José Betancourt is currently a PhD student at the Toulouse Mathematics Institute. His doctoral thesis is part of the ongoing ANR project RISCOPE, which focuses on the development of metamodeling techniques for coastal flooding early warning. His responsibility in the project lies in the frame of objective-based dimension reduction & meta-modeling for functional-input hydrodynamic codes.

MascotNum2019 conference - Abstract submission

T. BITTAR
CERMICS, Ecole des Ponts ParisTech / PRISME Department, EDF R&D

Supervisor(s): J-Ph. Chancelier (Ecole des Ponts ParisTech) and J. Lonchampt (EDF R&D)

Ph.D. expected duration: Feb. 2018 - Jan. 2021

Adress: 6 quai Watier, 78400 Chatou, Batiment S, Bureau S211

Email: thomas.bittar@edf.fr

Abstract:

EDF (*Electricité De France*) is a world leader in electricity generation and manages a large number of industrial assets. An industrial asset can be any physical installation managed by the company (a hydroelectric plant for example). In order to ensure effective and reliable electricity generation, the exploitation of these installations must be optimized. This task falls into the field of industrial asset management. For a given asset, several management strategies can be defined. More precisely, we focus on maintenance strategies that consist in setting up rules for the maintenance of components of a physical system. These strategies represent investments for the company. The goal of asset management is to provide indicators for decision support taking into account all technical and economic dimensions throughout the life of an asset, ensuring that the best investments are done are the right time.

In this work we consider a physical system with components sharing a common stock of spare parts. The performance of a maintenance strategy is quantified with an economic indicator: the NPV (*Net Present Value*). The NPV is the difference between the cost generated by a reference maintenance strategy and the evaluated maintenance strategy. Hence a positive NPV means that the evaluated strategy is better than the reference one. As the costs depend on random failures that can happen during the life of a component, the NPV is itself a random variable. EDF has developed the software VME (*Valorisation Maintenance Exceptionnelle*) that uses Monte-Carlo simulations to estimate the distribution of the NPV. We denote by $j(u, \omega)$ the output of one simulation, which is the NPV for the maintenance strategy u with the realization ω of the random variable W modelling the dates of failure of the components. The performance of a given maintenance strategy can then be quantified by computing an estimation of the risk measure we consider on the NPV (expectation, α-quantile, ...). Here, the risk measure we use is the expectation $\mathbb{E}(j(u, W))$, i.e. the best maintenance strategy is the one which leads to the highest expectation for the NPV.

VME is therefore able to evaluate the performance of a given maintenance strategy, however it is not possible to do optimization with the software in order to find the best maintenance strategy

$$u^\star \in \arg\min_{u \in U} J(u) \text{ where } J(u) = \mathbb{E}(j(u, W)) \tag{1}$$

where U is the set of admissible maintenance strategies. Solving this optimization problem is the goal of this work.

The code VME is considered to be a blackbox: given an input maintenance strategy u it outputs an estimation $\hat{J}(u)$ of $J(u)$ but we have no access to the gradients of J, they are in fact not even necessarily defined. Moreover the evaluation of the objective function J is noisy as the Monte-Carlo method only gives estimations of the expectation. Finally, for large systems, we also need to take into account that one function evaluation, i.e. the computation of $\hat{J}(u)$ for one given $u \in U$ is expensive in computation time. This is due to the fact that we need a large number of Monte-Carlo simulations for one evaluation of J as the variance of the NPV $j(u, W)$ is large.

First, we compare two optimization techniques that are adapted to this framework, namely EGO (*Efficient Global Optimization*) [4] based on kriging techniques and a direct search technique called MADS (*Mesh Adaptive Direct Search*) [1]. At each iteration, EGO uses a metamodel of the code to evaluate the objective function at a promising point for optimization and updates the metamodel accordingly. MADS looks for evaluation points on a mesh which is updated at each iteration depending on the outcome of the evaluation. The rules for updating the mesh and choosing the directions of the search for evaluation points ensure convergence to a local minimum. These two algorithms are compared on the COCO (*COmparing Continuous Optimizers*) platform [3], which is a benchmark designed to assess performance of blackbox optimization algorithms. MADS turns out to be more efficient than EGO on this benchmark.

However, neither MADS nor EGO can tackle directly the optimization of maintenance for large systems as the dimension of the problem becomes too large. A methodology based on a decomposition-coordination method and the more general framework of the auxiliary problem principle [2] is then proposed to split the global optimization problem into subproblems of smaller size. Typically, we want that a subproblem only involves optimization with respect to one component only or the stock. An iterative algorithm consisting of the resolution of the subproblems followed by a coordination step between all local solutions is given. The appeal of this technique is that each subproblem can be solved by any method. Here the low-dimension of the subproblems will make them adapted to the use of MADS. The subproblems can also be solved in parallel as they are independent thus reducing the computation time.

Numerical tests are yet to be performed on a toy system of two components sharing a common stock of spare parts. If successful on this small case, the optimization for the case of a large number of components should also be tractable as it just results in more small subproblems to be solved at each iteration.

References

[1] Charles Audet and J. E. Dennis, Jr. Mesh Adaptive Direct Search Algorithms for Constrained Optimization. *SIAM Journal on Optimization*, 17(1):188–217, 2006.

[2] Pierre Carpentier and Guy Cohen. *Dcomposition-coordination en optimisation dterministe et stochastique*, volume 81 of *Mathmatiques et Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017.

[3] Nikolaus Hansen, Steffen Finck, and Raymond Ros. Coco-Comparing Continuous Optimizers: The documentation. Technical Report RT-0409, INRIA, 2011.

[4] 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.

Short biography – Before starting my PhD I obtained a degree in general engineering from Ecole Centrale de Lyon and a Master in Applied Mathematics from the University of Cambridge. My PhD is funded by EDF as part of the AMPH project (Asset Management for Hydraulics) whose goal is to increase the reliability and performance of the hydroelectric fleet. The main aspects of the project cover the evaluation of the risk of failures of materials and the determination of robust maintenance poclicies. My PhD fits exactly in the latter category as it aims at optimizing maintenance for systems with a large number of components.

Improvement of error covariance matrix computation in variational methods

SIBO CHENG
EDF R&D
LIMSI, CNRS, Univ.Paris-Sud, Université Paris-Saclay, F-91405 Orsay

Supervisor(s): Dr. Didier Lucor (LIMSI, CNRS), Dr. Jean-Philippe Argaud (EDF R&D), Dr. Bertrand IOOSS (EDF R&D) and Dr. Angélique Ponçot(EDF R&D)

Ph.D. expected duration: Dec. 2017 - Dec. 2020

Adress: 7 Boulevard Gaspard Monge, 91120 Palaiseau

Email: sibo@limsi.fr

Abstract:

The idea of variational data assimilation methods (e.g. *3D-VAR*, *4D-VAR*) consists of finding a compromise between background predictions and instrumental observations where the associated weights are provided by prior error covariance matrices. A key element is the improvement of background error covariance matrix (often denoted by B), giving the constraint of shortage of experimental information. The mis-specification of background error covariance matrix structure can be problematic in terms of reconstruction/prediction accuracy as well as output error covariance estimation. Continuous attention and effort has been given to this topic, several methods are developed in order to improve the B matrix computation. Considered as the main contributor of background error in a dynamical data assimilation chain, the model error covariance matrix computation is also carefully studied, as described in the overview [2]. These methods we find in literature are more appropriate in a successive data assimilation procedure while for our applications, we are also interested in short term prediction and statistic reconstruction.

A great effort has also been carried out to diagnose and improve the covariance matrices modelling *a posteriori*, in particular the diagnostic and iterative methods developed by Météo-France [1] (also known as Desroziers iterative method). This method adjusts sequentially the background-observation error covariance ratios based on posterior indicators. Recent efforts are also investigated to apply diagnostic methods in local sub-spaces, which could make the covariance rebuild more flexible especially for high dimensional or multivariate problems. However, we notice that the Desroziers iterative method only modifies a multiplicative coefficient of matrix B which means the assumed prior error correlation can not be corrected. The efficiency of this method could also be limited when lack of historical data due to sampling errors when evaluating posterior indicators.

Inspired by existing industrial practice, consisting in repeating several times the assimilation procedure with the same observations, we have developed two novel iterative methods: **CUTE** (Covariance Updating iTerative mEthod) and **PUB** (Partially Updating BLUE method) for building background error covariance matrices in order to improve the assimilation result under the assumption of a good knowledge of the observation error covariances.

Using a linear observation operator, we have compared *CUTE*, *PUB* with the Desroziers approach, starting by a mis-specified assumed background matrix B_{A} in a twin fluid mechanics experiment framework together. The improvement in terms of assimilation accuracy is similar for all three methods. However, experiments show that the two new methods own a significant advantage concerning output correlation recognition under the assumption that the background error is dominant over the one of the observations. We draw the evolution of reconstruction error against the number of iterations in Fig.1 [left].

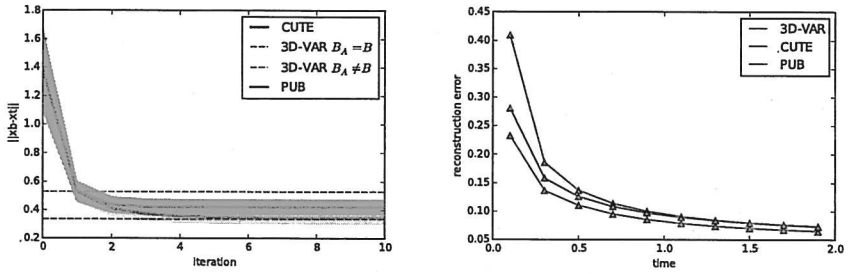


Figure 1: Comparison of standard *3D-VAR* method with *CUTE* and *PUB* in twin experiment framework in both a static reconstruction [left] and a successive data assimilation chain [right], starting with a mis-specified matrix B_{A}. For the static reconstruction, we draw the evolution of assimilation error and the standard deviation associated(transparent zones) throughout *CUTE*, *PUB* iterations, comparing to the one-shot *3D-VAR* algorithm when the background matrix is provided ($B = B_{\mathrm{A}}$, considered as the optimal target) or not ($B \neq B_{\mathrm{A}}$) . In the successive process, *CUTE* and *PUB* are only applied at the beginning (first reconstruction) of the process, following by standard *3D-VAR* algorithm latter. The initial background error is set to be 100 times higher than the observation error.

In a successive data assimilation chain, significant improvement provided by these two novel methods has also been identified compare to the flow-independent *3D-VAR* method, especially for short-term prediction. This advantage can be kept longer in the dynamical process as shown in Fig.1 [right] if the assumption of high level background/model noise is well fulfilled.

In order to get a more careful diagnostic, our effort has also been given to separate the state space into several well chosen sub-spaces where posterior diagnostic could be carried out independently. Instead of spatial distance based segmentation (as in [3]), we make the space separation by an observation based connection network. Unsupervised graph based community detection algorithms are therefore considered helpful for state space separation. For instance, reasonable positive results have been found in twin experiments by applying Desroziers iterative method in spatial distance independent sub-spaces using artificially simulated transformation operators.

Future focus will be given on the performance of our newly developed methods (*CUTE*, *PUB* as well as observation based connection networks) in more realistic/sophisticated industrial models.

References

[1] G. Desroziers and S. Ivanov. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, 127:1433 – 1452, 04 2001.

[2] P. Tandeo, P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido, and Y. Zhen. Joint estimation of model and observation error covariance matrices in data assimilation: a review. *Arxiv, accepted for submission to Monthly Weather Review*, 07 2018.

[3] J. A. Waller, S. L. Dance, and N. K. Nichols. On diagnosing observation-error statistics with local ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(708):2677–2686, 2017.

Short biography – Graduated from a master in applied mathematics, I started a PhD in 2017 at UPSud under a CIFRE convention with EDF R&D, where I completed an internship previously. The present subject is motivated by industrial problems in data assimilation reconstruction confronted by EDF.

Chance constraint optimization of a complex system - Application to the design of a floating offshore wind turbine

ALEXIS COUSIN
IFPEN & Ecole Polytechnique

Supervisor(s): Prof. Josselin Garnier (Ecole Polytechnique), Dr. Miguel Munoz-Zuniga (IFPEN), Dr. Martin Guiton (IFPEN) and Dr. Yann Poirette (IFPEN)

Ph.D. expected duration: Sep. 2018 - Sep. 2021

Adress: IFP Energies nouvelles, 1-4 avenue de Bois-Préau, 92500 Rueil-Malmaison, France

Email: alexis.cousin@ifpen.fr

Abstract:

The problem under interest in this PhD is a Reliability-Based Design Optimization (RBDO) [1] applied to the design of a component of a floating offshore wind turbine. Indeed, the reliability of this structure is in particular insured by the anchoring system of the floating support which restricts the wind turbine motion. This mooring system must have an attractive cost and avoid the ruin caused by a failure of the anchoring lines as a consequence of accumulated damage during the lifespan of the structure. We introduce randomness in the problem by considering uncertainties on the modeling process, represented by the random vector ξ, and on the marine environment conditions, represented by the random process Z. This problem can be mathematically stated as

$$\begin{aligned} \min_{x\in\Omega} \quad & c(x) \\ \text{s.t} \quad & \mathbb{P}(g(x,\xi) > \rho) < 10^{-4} \\ & \mathbb{P}(\min_{t\in[0,T]} \mathcal{T}(t,x,\xi;Z) < 0) < 10^{-4} \\ & \mathbb{P}(\max_{t\in[0,T]} \beta(t,x,\xi;Z) > 6) < 10^{-4} \end{aligned}$$

with :

- x the design variables and $\Omega \subset \mathbb{R}^n$ the feasible set ;
- c the cost of the mooring system ;
- g the fatigue constraint. g has strong non-linearities and is expensive to evaluate ;
- $\mathcal{T}$ the tension of the mooring line ;
- β the constraint on the pitch of the floater (rotation around the vertical axis) that must be less than 6° ;
- $Z = (Z(t))_{t\in[0,T]}$ represents the sea elevation and is a locally stationary process that is piecewise stationary over intervals of duration ΔT : $Z(t) = \sum_i \mathbb{1}_{[T_i,T_{i+1}]}(t)\zeta_i(t-T_i)$ where $T_i = i\Delta T$, $(\zeta_i(t))_{t\in[0,\Delta T]}$ are independent stationary Gaussian processes with covariance parameterized by $X_{LT,i}$. $X_{LT,i}$ are i.i.d. random vectors with given discrete joint probability distribution that characterize each stationary sea state.

The calculation of c does not raise any issue. Apart from the judicious choice of the optimization method to obtain an acceptable solution, the difficulty lies in the need to estimate a probability

in the optimization loop. This estimation cost comes from the characteristics of the constraints and the high level of acceptance probability. A naive approach by Crude Monte-Carlo is then prohibited.

We propose a methodology that takes into account the nature of the constraints to reduce the calculation cost.

For the fatigue constraint, there are already methods to deal with the calculation of this type of probability. Methods based on design point (FORM, SORM) require the solving of an optimization problem. Thereby, the RBDO problem becomes a double loop optimization. Different approaches as SORA [2] and SAP [3] have been considered to switch to a single loop optimization problem. We will apply, in a first time, these approaches which will serve as reference results. These methods reduce the computational effort but can lead to incorrect estimate of the probability of failure. Thus, in a second step, we will focus on solving the optimization with the contribution of metamodels coupled with improved Monte-carlo methods [4, 5].

The second and third constraints are evaluated thanks to the Extreme Value Theory [6] to make efficient approximations. Indeed, under some assumptions over a stationary Gaussian process ζ, we have the following result :

$$\mathbb{P}\left(a_T\left(\max_{t\in[0,T]}\zeta(t) - b_T\right) \leq \alpha\right) \to \exp(-e^{-\alpha}) \quad \text{as} \quad T \to \infty$$

where a_T depends on T and b_T depends on T and on the second spectral moment of ζ.

The next step will be to add another constraint for the extreme response design which will bring new difficulties due to the high dimension of the uncertainty vector involved.

The methodology is applied to the floating offshore wind turbine used in the benchmark OC4 [7].

References

[1] Nicolas Lelièvre, Pierre Beaurepaire, Cécile Mattrand, Nicolas Gayton, and Abdelkader Otsmane. On the consideration of uncertainty in design: optimization - reliability - robustness. *Structural and Multidisciplinary Optimization*, 54(6):1423–1437, Dec 2016.

[2] X. Du and W. Chen. Sequential Optimization and Reliability Assessment method for Efficient Probabilistic Design. *ASME J. Mech. Des.*, 126(2):225–233, 2004.

[3] G.D. Cheng, L. Xu, and L. Jiang. Sequential approximate programming strategy for reliability-based optimization. *Computers and Structures*, 84(21):1353–67, 2006.

[4] V. Dubourg. *Adaptive surrogate models for reliability analysis and reliability-based design optimization.* PhD thesis, Université Blaise Pascal, Clermont-Ferrand, France, 2016.

[5] Javiera Barrera, Tito Homem-De-Mello, Eduardo Moreno, Bernardo K. Pagnoncelli, and Gianpiero Canessa. Chance-constrained Problems and Rare Events: An Importance Sampling Approach. *Math. Program.*, 157(1):153–189, May 2016.

[6] M.R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes.* Springer series in statistics. Springer-Verlag, 1983.

[7] A. Robertson, J. Jonkman, M. Masciola, H. Song, A. Goupee, A. Coulling, and C. Luan. Definition of the Semisubmersible Floating System for Phase II of OC4. *NREL Technical report NREL/TP-5000-60601*, 2012.

Short biography – Alexis Cousin got a Master's Degree from University of Paris-Saclay in modeling and numerical simulation. He is currently a first year PhD student. This thesis, funded by IFPEN, focuses on a Reliability-Based Design Optimization problem applied to the design of an offshore wind turbine.

Clustering multivariate functional data defined on random domains: an application to vehicle trajectories analysis

S. GOLOVKINE
National School for Statistic and Information Analysis (ENSAI)

Supervisor(s): Prof. Patilea (ENSAI), Prof. Klutchnikoff (University Rennes 2) and Dr. Cembrzynski (Renault)

Ph.D. expected duration: Jan. 2018 - Dec. 2020

Adress: 3, rue Philibert Delorme, 78280 Guyancourt, FRANCE

Email: steven.s.golovkine@renault.com

Abstract:

With the recent development of sensing devices, more and more data are recorded continously (or at least at high frequency) through time and space. These measures lead to large amount of data, called *functional data*. We retrieve functional data in large variety of domains. For example, in biology, growth curves have probably been the first dataset considered as functional data [3]. But, one can find application in physics (spectroscopy), economics (index evolution), musics (sounds recognition), medicine (electroencephalography comparison), and so on. Lately, multivariate functional data have been considered. For instance, one can cite two famous examples from Ramsey and Silverman [2], gait cycle data and Canadian weather data. Moreover, the automotive industry also generates large volume of functional data. In particular, vehicle trajectories, which are our case of interest, could be describe like that. Functional data analysis (FDA) develops the theory and statistical methodology for studying such data. So, FDA is the analysis of data that are, in a general manner, objects that can be represented by functions. Thus, by analogy with multivariate data analysis where an observation is represented by a random vector of scalars, in FDA, an observation is a random vector of functions. Hence, functional data are intrinsically infinite dimensional. However, we can not generally observed directly the functions but only a discretization of the functions over a fxed or random grid of points.

Now, recall our case of interest: vehicle trajectories. Nowadays, a vehicle records a lot of information about its environment through his different sensors (camera, radar, lidar). More particularly, it registers some characteristics about vehicles around him at high frequency. These characteristics can be the longitudinal and lateral position, the acceleration, the size, the type of the vehicle for instance. All the information are recorded relatively to the considered vehicle (also known as EGO car). Define a driving scene as a small period of time, say $\mathcal{T}$, during which we record the environment of the EGO car. This environment is constituted by a certain number of vehicles, say P, whose one records a certain number of characteristics for each vehicle, say D. However, we do not assume that all of the P vehicles are recorded on the complete interval $\mathcal{T}$, but only on a random compact subset of $\mathcal{T}$. So, an observation of a scene can be represented as a random vector of functions :

$$\mathbf{Z} = \left(Z^{(1)}, \ldots, Z^{(P)}\right), \quad \text{where} \quad \forall i \in [\![1, P]\!], Z^{(i)} : \mathcal{T}^{(i)} \subset \mathcal{T} \longrightarrow \mathbb{R}^D.$$

Moreover, $Z^{(i)}$ is assumed to be in $L^2(\mathcal{T}^{(i)})$ for all $i \in [\![1, P]\!]$. So, realizations of $\mathbf{Z}$ are multivariate functional data which are defined on different domains. The analysis of such data is performed in three major steps: smoothing, dimension reduction and then clustering.

The smoothing step has two major goals. The first one is to remove the eventual noise in the measurements because the sensors are not perfect and they can not retrieve exactly the reality.

Secondly, as the functions are defined on different domains, we use change-of-methods to put them on a common interval, for instance $[0, 1]$. Here, the smoothing is performed to resampled the functions on a common grid such that they will be comparable.

In a second time, dimension reduction is done using multivariate functional principal components analysis [1]. The idea is to write all the observations of the scenes into a common multivariate basis of functions. In fact, the multivariate version of the Karhunen-Loève decomposition told us that:

$$\mathbf{Z}(t) = \boldsymbol{\mu}(t) + \sum_{j=1}^{\infty} c_j \boldsymbol{\Phi}_j(t),$$

where $\boldsymbol{\mu} = \left(\mathbb{E}(Z^{(1)}), \ldots, \mathbb{E}(Z^{(P)})\right)$ is the mean vector of each function, $\{\boldsymbol{\Phi}_j\}_{j\geq 1}$ are the multivariate eigenfunctions found by an eigenanalysis of the covariance operator of $\mathbf{Z}$ and the c_j are the projection of $\mathbf{Z}$ onto $\boldsymbol{\Phi}_j$. In practice, we truncate the Karhunen-Loève expansion at M terms. This truncation is the optimal approximation of $\mathbf{Z}$ of dimension M. Usually, M is chosen to explain a certain percentage of variance (95% or 99% generally) of the data.

So, our multivariate functions $\mathbf{Z}$ are summarized by M coefficients. And the clustering step follows directly from that. Classical clustering algorithms are launched on the set of coefficient with a particular metric which take into account the variability of the data in the coefficients.

An algorithm is proposed to analyze vehicle trajectories data using such a methodology.

References

[1] C. Happ and S. Greven. Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, 113(522):649–659, April 2018. arXiv: 1509.02029.

[2] James Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer Series in Statistics. Springer-Verlag, New York, 2 edition.

[3] R.D. Tuddenham and M.M. Snyder. *Physical Growth of California Boys and Girls from Birth to Eighteen Years.* Publications in child development. University of California Press, 1954.

Short biography – I have a MSc in Big Data and an engineering degree in statistics from ENSAI. At the end of my end-of-study internship at Renault, I was proposed a thesis about the data analysis coming from the autonomous vehicle. So, I started as a PhD student at the beginning of January 2018 thanks to the CIFRE plan.

Polynomial chaos expansion for wave propagation

A. GOUPY
CEA & ENS Paris-Saclay

Supervisor(s): D. Lucor (LIMSI, CNRS, Université Paris-Saclay), C. Millet (CEA, DAM, DIF)

Ph.D. expected duration: Jan. 2017 - Dec. 2019

Adress: CMLA, ENS Cachan, CNRS, Université Paris-Saclay, 94235 Cachan, France

Email: alexandre.goupy@cmla.ens-cachan.fr

Wave propagation in a random medium

The problem of wave propagation through a random medium arises naturally in many physical applications. Many theoretical works have been done on this subject but its numerical simulation requires a strategy to overcome the numerical cost limitation. In fact, a naive approach would be far too expensive and the construction of a metamodel often falter over the long term integration problem ([2]).

Moreover, in the case of wave propagation, the received signal results from the superposition of interfering wave packets, each one depending on the stochastic characteristics of the medium. To deal with the complexity of the resulting signal, we propose a method to build a metamodel based on a decomposition adapted to the medium: the normal modes of the propagating operator.

A stochastic basis adapted to the propagation

After a Fourier transform in time, the problem of wave propagation results in solving the Helmoltz equation:

$$\mathcal{H}(x,\xi)u = \Delta u + \frac{\omega^2}{c(x,\xi)^2}u = s(\omega) \tag{1}$$

where $s(\omega)$ is the spectrum of the source and $c(x,\xi)$ the wave celerity in the medium. The randomness of the medium gives a wave celerity which depends on random paramaters ξ.
Linear operator theory ensures that the eigenvectors $(\Psi_k)_{k\in K}$ of $\mathcal{H}$ form a basis of the space of squarely integrable functions. This basis gives a natural decomposition in wave packets for the solution u.
However, this basis is defined as the spectrum of a random operator $(\mathcal{H}(x,\xi))$ and depends on the stochastic parameters ξ. We propose to consider the Polynomial Chaos expansion (gPC) of this basis in order to be able to decompose the solution for every realisation of our medium with a low computational cost.

A modular metamodel

Once the gPC expansions of the normal modes $(\widehat{\lambda_k}(\omega,\xi))_{k\in K}$ and $(\widehat{\Psi_k}(x,\omega,\xi))_{k\in K}$ are computed, they can be used to generate signals for a given source at a distance R:

$$u(\omega,R,\xi) = \left[\frac{i}{4}\sum_{k\in K} H_0^{(1)}(\widehat{\lambda_k}(\omega,\xi)R)\widehat{\Psi_k}^2(0,\omega,\xi)\right] s(\omega) \tag{2}$$

Since the metamodel is build upstream, a stochastic source can be considered without supplementary cost. Sensitivity analysis can also be conducted using those expansions.

Moreover, this appraoch gives a natural framework for model reduction: the sum can focus on the most contributing modes ([1]). For instance, by taking only one mode we have a metamodel for one wave packet which can be usefull when studying a particular arrival in a received signal.

Towards a multi-level approach

The wide range of possible perturbations can alter the convergence of the gPC expansion. We propose to seperate two scales of the perturbation: the large scales structures which are treated as described above and small-scale structures treated with a perturbative approach.

This perturbative approach relies on the coupling matrix between the acoustic modes. This coupling matrix depends on the large scale structures but its gPC expansion can be deduced from the expansions $(\widehat{\lambda_k}, \widehat{\Psi_k})_{k\in K}$.

References

[1] Michael Bertin, Christophe Millet, and Daniel Bouche. A low-order reduced model for the long range propagation of infrasounds in the atmosphere. *The Journal of the Acoustical Society of America*, 136(1):37–52, 2014.

[2] Xiaoliang Wan and George Karniadakis. Long-term behavior of polynomial chaos in stochastic flow simulations. *Computer Methods in Applied Mechanics and Engineering*, 195:5582–5596, 08 2006.

Short biography – This PhD has started in January 2017 at ENS Paris-Saclay and is funded by CEA. CEA is working with the CTBTO (Comprehensive Nuclear Test Ban Treaty Organization) on the detection of explosions on the surface of the globe using infrasound monitoring stations. The aim of this work is to take into account the impact of the atmospheric unertainties on infrasound propagation.

MascotNum2019 conference - Principal Component Analysis and "boosted" weighted least-squares method for training tree tensor networks

CECILE HABERSTICH
Ecole Centrale de Nantes

Supervisor(s): Anthony Nouy (Ecole Centrale de Nantes), Guillaume Perrin (CEA/DAM/DIF, F-91297, Arpajon)

Ph.D. expected duration: 2017 - 2020

Adress: CEA/DAM/DIF, F-91297, Arpajon

Email: cecile.haberstich@ec-nantes.fr

Abstract: One of the most challenging tasks in computational science is the approximation of high dimensional functions. Most of the time, only a few information on the functions is available, and approximating high-dimensional functions requires exploiting low-dimensional structures of these functions.
In this work, the approximation of a function u is built using point evaluations of the function, where the evaluations are selected adaptively. Such problems are encountered when the function represents the output of a black-box computer code, a system or a physical experiment for a given value of a set of input variables.
A multivariate function $u(x_1, ..., x_d)$ defined on a product set $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_d$ can be identified with a tensor of order d.
Here, we present an algorithm for the construction of an approximation of a function u in tree-based tensor format (tree tensor networks whose graphs are dimension partition trees). A low-order tensor v_α, seen as a vector-valued map, is associated to each node α of the dimension partition tree T, and this set of tensors totally parameterizes the approximation.

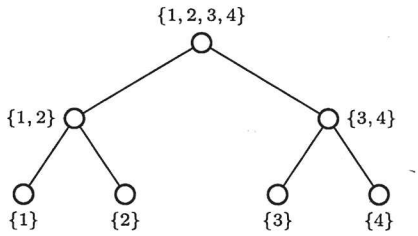


Figure 1: Example of a dimension partition tree T over $D = \{1, 2, 3, 4\}$

For example, an approximation v associated with the tree of the Figure 1 takes the form:

$$v = v_{1,2,3,4}(v_{1,2}(v_1(\phi_1(x_1)), v_2(\phi_2(x_2)), v_{3,4}(v_3(\phi_3(x_3)), v_4(\phi_4(x_4)))))$$

where the $\phi_\nu : \mathcal{X}_\nu \to \mathbb{R}^{n_\nu}, \nu \in \{1, 2, 3, 4\}$ are the feature maps.
The algorithm relies on an extension of principal component analysis (PCA) to multivariate functions in order to estimate the tensors. In practice, PCA is realized on sample-based projections of the function u, using interpolation or least-squares regression.
To provide a stable projection, least-squares regression usually requires a high number of evaluations of u, which is not affordable in our context. This number of evaluations can be decreased thanks to a so-called "boosted" weighted least-squares method. This method combines an optimal weighted least-squares method proposed in [1] and a re-sampling technique. With a particular choice of weights and samples and through re-sampling, an approximation error of the order of

the best approximation error is guaranteed using a moderate number of samples, of the order of the dimension of the approximation space.
We use this methodology in our algorithm and will compare it with strategies using standard least-squares method or interpolation (as proposed in [2]).

References

[1] A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *SMAI Journal of Computational Mathematics*, 3:181203, 2017.

[2] A. Nouy. Higher-order principal component analysis for the approximation of tensors in tree-based low rank formats. *Numerische Mathematik*, 2019.

Short biography – I graduated from Centrale Nantes with a specialization in applied mathematics and from the Technical University of Munich with a specialization in computational mechanics in 2017. I began a PhD thesis whose subject is "Low-rank approximation methods for complex uncertainty quantification problems". This thesis is a joint work between Centrale Nantes and the CEA DAM.

Sensitivity analysis of an avalanche flow dynamics model using aggregated indices

MARÍA-BELÉN HEREDIA
Irstea, Université Grenoble Alpes

Supervisor(s): Nicolas Eckert (Irstea) and Clémentine Prieur (Université Grenoble Alpes)

Ph.D. expected duration: Oct. 2017 - Sep. 2020

Adress: Irstea, 12 Rue de la Papeterie, Grenoble

Email: maria-belen.heredia@irstea.fr

Abstract:

Avalanche flow dynamics models depend on inputs that are poorly known (e.g. friction parameters, initial conditions corresponding to the avalanche release, etc). The outputs of these models are commonly both functional and scalar and they are employed for land-use planning and the design of defense structures. Thus, it is required to assess the impact of the uncertainty of the parameters on the outputs, and this is the aim of sensitivity analysis. It is possible to apply sensitivity analysis to each output of the model separately but this leads to redundancy in the results. An alternative based, on aggregated Sobol' indices was proposed by [2] (see also [1]). We propose here to reduce functional outputs to vectorial ones and then to compute aggregated Sobol' indices. Specifically, we developed the sensitivity analysis of two functional and one scalar output of an avalanche dynamics model. First, the outputs are decomposed in basis functions using simultaneous principal components and then, the generalized Sobol' sensitivity indices are computed on the coefficients of the expansion. Application is made to a fluid avalanche model based on depth-averaged Saint-Venant equations on a typical avalanche path. The results show that the Coulombian friction coefficient is the most influential input of the model on a case study path but the influence of the other inputs is not negligible.

References

[1] F. Gamboa, A. Janon, T. Klein, and A. Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *ArXiv e-prints*, November 2013.

[2] Matieyendou Lamboni, David Makowski, Simon Lehuger, Benoit Gabrielle, and Herv Monod. Multivariate global sensitivity analysis for dynamic crop models. *Field Crops Research*, 113(3):312 – 320, 2009.

Short biography – I'm a second year PhD student. I have an engineering Mathematics degree from the Escuela Politécnica Nacional (Ecuador) and a master's degree in Applied Mathematics from the University of Grenoble. My doctoral project is about the sensitivity analysis and the Bayesian calibration of avalanche models using data of high spatio-temporal resolution. This project is founded by OSUG@2020 and the CDP-Trajectories framework.

Maximum Entropy on the Mean approach to solve inverse problems with an application in computational thermodynamics.

EVA LAWRENCE
Université Paris Saclay - CEA DEN SCCME

Supervisor(s): Fabrice GAMBOA (Institut de Mathématiques de Toulouse), Christine GUENEAU (CEA) and Thierry KLEIN (ENAC)

Ph.D. expected duration: Oct. 2017 - Oct. 2020

Address: Université Paul Sabatier, Institut de Mathématiques de Toulouse,
118 route de Narbonne, 31062 TOULOUSE Cedex 9

Email: eva.lawrence@math.univ-toulouse.fr

Abstract:

In the context of computational thermodynamics, we aim at reconstructing a multidimensional function $f : \mathbb{R}^d \to \mathbb{R}^p$ which is solution of an inverse problem. That is, knowing a training set $\{x_l, z_l\}_{l=1,\dots,N}$, we aim at building a regularized $\mathbb{R}^p$-valued function $f = \begin{pmatrix} f^1 & \dots & f^p \end{pmatrix}$ such that

$$\sum_{i=1}^{p} \lambda^i(x_l) f^i(x_l) = z_l, \qquad l = 1, \dots, N, \tag{1}$$

with given functions λ^i. Component f^i represents an energy function. We propose a regularized solution for this problem by a Maximum Entropy on the Mean (MEM) method. Interpolation problem, as problem stated in (1), defines a too "local" constraint and will lead to a trivial reconstruction by MEM methods. Mixed interpolation and moment constraints must be considered to find an appropriate solution.

Motivated by crystallography [1], MEM method has been developed in [2] and [3]. The method aims at the reconstruction on space U of the probability measure P associated with random variables Y when having at hand only a few information on Y.

Let P_0 be a probability measure defined on compact space U. P_0 will be called the reference measure. MEM method derives as solution P the probability measure with highest P_0-entropy, that is the probability measure P which minimizes the divergence (or maximizes the entropy) from measure P_0. Reference measure P_0 can act as a priori information on Y distribution. Letting K be the Kullback-Leibler divergence, this problem is more formally written

$$\begin{aligned} &\min\ K(P, P_0) \\ &P\ :\ \int_U y_l dP(y) = z_l, \quad \cdot\ z = \begin{pmatrix} z_1 & \dots & z_N \end{pmatrix}^T \in \mathbb{R}^N. \end{aligned} \tag{2}$$

To estimate the solution P, we will work on a sequence of estimators

$$\nu_n = \frac{1}{n} \sum_{i=1}^{n} X_i \delta_{t_i}, \tag{3}$$

where X_i are random amplitudes and δ_{t_i} is the discretization of space U.

In [4] the authors have proposed an extension of MEM method for functional reconstruction instead of probability measure reconstruction. The key idea of MEM method in this case is that

the function to be reconstructed is seen as the expectation of a random function with respect to the unknown probability measure P on the space of functions. The idea is to link a problem in convex optimisation with problem (2).

In convex analysis, the selection of a multidimensional function f can be proceeded by a convex minimization problem. Given a certain convex function γ

$$\begin{aligned} &\min \int_U \gamma(f^1, \ldots, f^p) dP_U \\ &f \ : \ \int_U \sum_{i=1}^{p} \lambda^i(x) f^i(x) \varphi_l(x) dP_U(x) = z_l \qquad 1 \leqslant l \leqslant N. \end{aligned} \tag{4}$$

By a linear transfer principle linking function f to some measure F, we will show that solving (4) can be brought back to solving the following problem on measures taking D_γ, the divergence built in line with convex function γ

$$\begin{aligned} &\min \ D_\gamma(F, F_0) = \min \int_U \gamma\left(\frac{dF^1}{dP_U}, \ldots, \frac{dF^p}{dP_U}\right) dP_U \\ &F \ : \ \int_U \sum_{i=1}^{p} \lambda^i(x) T\left(F^i(x)\right) d\Phi_l(x) = z_l \qquad 1 \leqslant l \leqslant N, \end{aligned} \tag{5}$$

with T is transfer operator from the measure space to the function space. We can go back to a finite dimensional problem by a discrete approximation of F amplitudes. MEM estimator is then obtained as the limit of the discretized estimators for the finite-dimensional problem.

As an energy function, component f^i in the context of thermodynamics must reflect some physical properties and so, it is expected to be regular in a certain sense. Choice of reference measure P_U leads to different levels of regularity for the reconstructed functions f^i. The MEM method described will be applied to a toy example built in agreement with thermodynamics requirements.

References

[1] Jorge Navaza. On the maximum-entropy estimate of the electron density function. *Acta Crystallographica Section A*, 1985.

[2] Didier Dacunha-Castelle and Fabrice Gamboa. Maximum d'entropie et problème des moments. *Annales de l'I.H.P. Probabilités et statistiques*, 1990.

[3] Fabrice Gamboa and Elisabeth Gassiat. Maximum d'entropie et problème des moments: cas multidimensionnel. *Probability and Mathematical Statistics*, 1991.

[4] Imre Csiszár, Fabrice Gamboa, and Elisabeth Gassiat. MEM pixel correlated solutions for generalized moment and interpolation problems. *IEEE Transactions on Information Theory*, 1998.

Short biography – Eva Lawrence graduated from Université Paul Sabatier in 2017 with a Master Degree in Applied Mathematics. She started her PhD at Institut de Mathématiques de Toulouse funded by CEA Saclay in October 2017. PhD project deals with estimating a class of energy functionals in Thermodynamics, namely Gibbs free enthalpy, and studying in this frame uncertainty propagation.

Sparse polynomial chaos expansions: Benchmark of compressive sensing solvers and experimental design techniques

NORA LÜTHEN
ETH Zurich, Switzerland

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

Ph.D. expected duration: Apr. 2018 - Apr. 2022

Adress: Stefano-Franscini-Platz 5, 8093 Zurich

Email: luethen@ibk.baug.ethz.ch

Abstract: Polynomial chaos expansions (PCE) are a well-known and popular surrogate modelling technique that expands the model response in terms of orthogonal polynomial functions of the input random variables. While PCEs work well for low input dimensions, the accurate computation of their coefficients becomes challenging in high dimensions, because the number of basis functions (and hence coefficients) grows exponentially with the dimension. The same holds for the case when polynomials of high degree are required to achieve a good approximation. At the same time, traditional methods for computing the coefficients of a PCE, such as projection or least-squares regression, require a number of model evaluations that is larger than the number of basis functions. Both challenges limit the applicability of such PCE methods to high-dimensional, costly models.

Fortunately, these issues can be addressed by using the Compressive Sensing framework to compute a sparse PCE, i.e., one for which only few of the coefficients are nonzero. Here, the regression problem is modified by adding a constraint on the sparsity of the solution. Sparse PCEs perform well if the model is compressible, i.e., if the coefficients of a high-dimensional PC approximation to the model decay sufficiently fast. This is usually the case for real-world models. Moreover, sparse PCEs need by far less model evaluations than traditional methods, which enables their use in high-dimensional and high-degree settings.

In recent years, a large number of articles has been published that propose efficient methods for computing sparse PCEs from a small number of model evaluations, using ideas from Compressive Sensing. Many of these contributions have good theoretical guarantees as well as superior performance on example problems. However, comparisons are often only made with respect to standard methods, not to other recent developments. Also, the methods vary considerably in computational demand. For engineers who want to apply sparse PCE to their problems but not read and evaluate the large literature on the topic, a guideline is needed to decide which method shall be used in a given situation.

Our contribution is a literature review together with extensive numerical benchmarking. We collect and explain the available methods and analyse their behavior on various analytical and numerical examples. We also propose a general modular framework for adaptive sparse PCE computations, in which most of the methods put forward in the literature can be fit. In particular, the main modules are basis adaptivity, sampling or enrichment of the experimental design, and computing a solution to the sparse regression problem. The adaptive sparse PCE procedure consists of the repeated execution of these modules.

For each of the modules, many methods have been proposed in the literature. As an example, for creating the experimental design there are

- space-filling methods such as Sobol sequences and Latin hypercube sampling (LHS);

- random methods such as Monte Carlo sampling and coherence-optimal sampling [4]; and
- methods based on optimizing a scalar function of the PC evaluation matrix over a pool of candidate samples, such as D-optimal [2], S-optimal [3] and near-optimal [1] sampling.

Likewise, many methods have been proposed for the solution of the sparse regression problem. The clear structuring of the sparse PCE procedure into modules naturally leads to a few new combinations of methods that have not yet been considered in the literature. Numerical results on example problems are presented that help guide the decision about which method shall be used in which situation.

Figure 1 shows exemplary results for surrogating the Ishigami function with a sparse PCE, using a number of different ED sampling techniques and a basis of total degree 20 (with 1771 candidate regressors). LAR is used to solve the sparse regression problem.

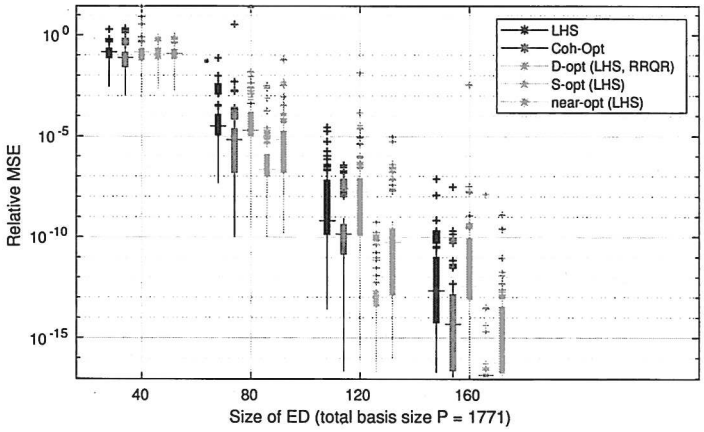


Figure 1: Surrogating the Ishigami function with sparse PCE: Plot of relative MSE against number of design points for different ED sampling techniques (sparse solver: LAR, basis: total degree ≤ 20, 100 replications).

References

[1] N. Alemazkoor and H. Meidani. A near-optimal sampling strategy for sparse recovery of polynomial chaos expansions. *J. Comp. Phys.*, 371:137–151, 2018.

[2] P. Diaz, A. Doostan, and J. Hampton. Sparse polynomial chaos expansions via compressed sensing and D-optimal design. *Comput. Methods Appl. Mech. Engrg.*, 336:640–666, 2018.

[3] N. Fajraoui, S. Marelli, and B. Sudret. Sequential design of experiment for sparse polynomial chaos expansions. *SIAM/ASA J. Unc. Quant.*, 5(1):1061–1085, 2017.

[4] J. Hampton and A. Doostan. Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies. *J. Comp. Phys.*, 280:363–386, 2015.

Short biography – Nora has studied mathematics with an emphasis on numerical mathematics at the University of Bonn, Germany. Since 2018, she is PhD student in Prof. Sudret's Chair of Risk, Safety and Uncertainty Quantification at ETH Zurich in Switzerland. Her PhD is part of the project "Surrogate modelling for stochastic simulators" funded by the Swiss National Science Foundation.

A predictive Data Driven Approach based on Reduced Order Models for the Morphodynamic Study of a Coastal Water Intake

Rem-Sophia Mouradi
EDF R&D LNHE (Laboratoire National d'Hydraulique et Environnement) and CERFACS (Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique), Université de Toulouse, INPT (Institut National Polytechnique de Toulouse)

Supervisor(s): Prof. Olivier Thual (Université de Toulouse, Institut de Mécanique des Fluides de Toulouse IMFT, INPT and CERFACS), Dr. Cédric Goeury (EDF R&D LNHE), Dr. Fabrice Zaoui (EDF R&D LNHE) and Dr. Pablo Tassi (EDF R&D LNHE and Saint-Venant Laboratory)

Ph.D. expected duration: Feb. 2018 - Feb. 2021

Adress: EDF R&D, 6 quai Watier, 78400 Chatou

Email: rem-sophia-r.mouradi@edf.fr

Abstract:

This work is motivated by the following question: How to deal with a dynamic physical problem when a numerical model is not an option (inexistent, unreliable or time consuming)? Specifically, when a fairly large data set is available, and the interest data are of different natures. In this context, we assess a dynamical data-driven model that links a two-dimensional state variable to scalar forcing variables.

The presented methodology is applied within the context of a power plant water intake monitoring. The intake channel is located in a coastal area, and must ensure enough water supply for the cooling process of the power plant, even though it is subject to massive sediment arrivals, which represents a clogging risk. One of the industrial challenges is therefore to predict the sediment dynamics observed in the channel, which can be deduced from the observed bed evolutions in the study area. The sediments outside of the channel can be stirred up under the waves constraint, and transported towards the channel by the tidal currents. This process can be amplified during low tide levels, resulting in a higher sediment volume entering the channel

Due to the monitoring needs, bathymetric measurements of the channel are performed on a regular basis, along with meteorological and hydrodynamic surveys (waves, wind, tidal levels, etc.) as well as management information (dredging data, pumping flowrates). The aim is therefore to establish a dynamical model that predicts the bed elevations state field, from the knowledge of the previous state and the several forcing parameters.

As the bed elevation is a two-dimensional field, it must be reduced to a representative vector of scalar variables by applying a Proper Orthogonal Decomposition (POD) [1]. The POD consists of decomposing a field that depends on both time and space variables into a finite sum of functions with separate variables. This allows first to isolate the spatial patterns, represented by the functions depending on space variables, called the POD basis. The POD basis terms, when added, explain the observed dynamic. The interest of this operation is that the deduced patterns can often be interpreted in terms of physical behavior. Another consequence of the POD is that the functions depending on time that are associated to each member of the POD basis, are simply scalar variables that vary in time. Their variability directly represents the variability of the original two-dimensional field, here the bathymetry. In a non-chaotic system, these functions, called "temporal modes", are often signals that can be explained by inputs, or even predicted by an adequate statistical model. This means that the temporal modes at future times can be predicted from previous times using the information on the forcing variables. A finite number **"K"** of modes can be selected as a reliable representation of reality.

A natural outgrowth of the analysis is therefore to propose a statistical model. In our study, we propose a Polynomial Chaos Expansion (PCE). The temporal modes are considered as random variables that can be linked to random forcing variables, considered as independent, via an orthogonal polynomial basis. The PCE is build using the Least Angle Regression Stagewise (LARS) method, which is a sparse PCE construction [2]. This allows to gain in fitting accuracy by increasing the polynomial degree with small amount of data.

The strength of the PCE is that the coefficients of the expansion are directly linked to the variance of the contribution of each variable and its interactions to the response. In other words, the Sobol sensitivity estimators deduced from the ANOVA decomposition can be calculated without further effort [3]. This allows to work both on the prediction model and on the quantification of correlation between the forcing parameters and the bathymetry response through its modes.

This analysis is done on each of the **"K"** chosen modes. The outcome of this step is therefore **"K"** dynamical prediction models linking each temporal mode to its future estimation.

The proposed prediction methodology is (see **Figure 1**):

1. Start from a new measured bathymetry: Project it on the constructed POD basis.
2. For each deduced temporal mode: predict the new value using the forcing variables and the previously constructed PCE models
3. Multiply each new temporal mode by its corresponding spatial mode (member of POD basis), and sum the all to reconstruct a new bathymetric state, that is physically consistent.

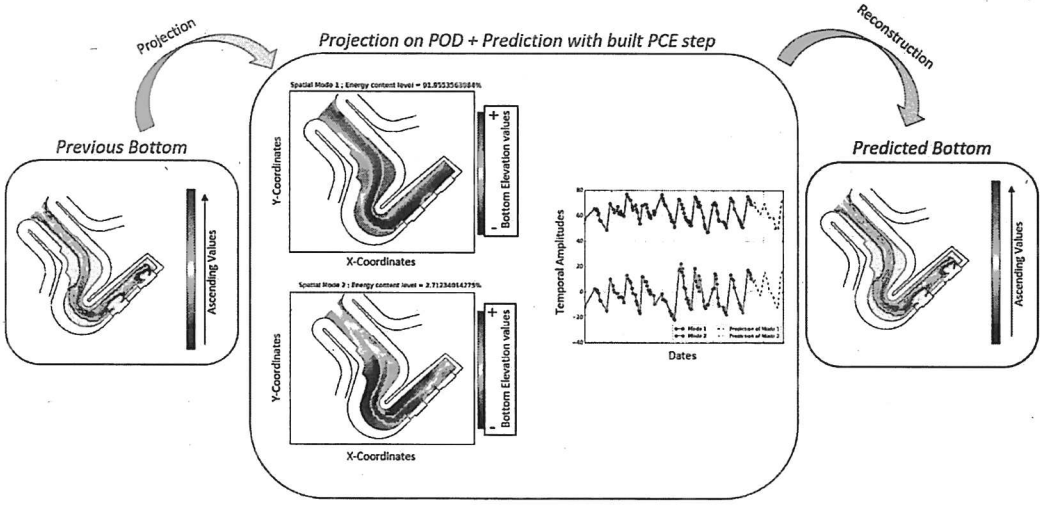


Figure 1 - Representation of the prediction algorithm

A number of uncertainty sources can be identified in this methodology, namely through the measurement errors, in the choice of the number of modes in the POD approximation, in the convergence of the POD itself, in the construction of the PCE and contribution of extreme events to the learning process, etc. This allows to characterize the prediction by a confidence interval and a density law for the residual errors.

As a perspective, the same procedure can be applied to evaluate the dynamics of a proposed process-based sediment transport model, and compare it to field observations. Furthermore, the decomposition of the numerical model response can make the calibration process (data assimilation) easier, by evaluating the temporal modes values instead of evaluating a complete two-dimensional field.

References

[1] Lumley, J. L. *The Structures of Inhomogeneous Turbulent Flow.* Proceedings of the International Colloquium on the Fine Scale Structure of the Atmosphere and Its Influence on Radio Wave Propagation, edited by A. M. Yaglam, and V. I. Tatarsky (Nauka, Moscow, 1967), pp. 166–178.

[2] Blatman, G., 2009. *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis.* Ph.D. thesis, IFMA/Université Blaise Pascal.

[3] Iooss, B. 2011. *Revue sur l'analyse de sensibilité globale de modèles numériques.* Journal De La Société Française De Statistique. 152.

Short biography – I studied mathematical and mechanical modeling, finishing on an internship about uncertainty quantification in sediment transport. My PhD results from the needs of a reliable prediction of bed movement in cooling intakes, which is one of EDF concerns, as dredging operations are costly. It is a PhD financed by ANRT and EDF R&D and directed in collaboration with CERFACS and INPT Université de Toulouse.

Extended Principal Component Analysis algorithm for adaptive model reduction in inverse problems

A. MUKHIN
Moscow Institute of Physics and Technology

Supervisor(s): Aleksey Khlyupin, PhD (MIPT, CET MIPT)

Ph.D. expected duration: Sep. 2019 - Jul. 2023

Adress: 4, Nauchny Lane, Dolgoprudny, Moscow area, 141700, Russia

Email: andrei.mukhin@phystech.edu

Abstract:

The problem of identifying model parameters given observed data appears in many areas of contemporary research. Such problems are called inverse problems and in most cases are ill-posed. The main approach for fast and efficient solution of this type of problems is to solve the corresponding optimization problem using variety of optimization techniques. Objective function in such optimization problems is typically written in form of squared misfit between observed data and model simulation results:

$$\min_x \left[F(x) = \sum_{i=1}^{N} \left\{ \frac{f(x_i) - d_i^{obs}}{\sigma_i} \right\}^2 \right] \tag{1}$$

where d_i^{obs} - observed data, x - model, $f(x_i)$ - function of model, σ_i - standard deviation.

In practical problems $f(x)$ is determined by physical processes modelling and has form of system of PDE's which is commonly packed into complex simulation systems or software products. Calculation of $f(x)$ is referred as solving forward problem. In practice, one run of the forward problem could take from few ms to few days of computation cost.

One of the main approaches to accelerate the convergence of an optimization algorithm for problem (1) is to parametrize model x or, in other words, to reduce model size. Typically this reduction is achieved by decomposition of the initial model using an orthogonal basis and further optimization in reduced space of decomposition parameters.

The model reduction problem has been widely studied in the last 30 years. In general, most of parametrization techniques can be classified into two groups [2]. First represents model decomposition as linear combination of fixed basis functions (e.g. Discrete Cosine Transform, Discrete Wavelet Transform), while the second group allows determination of the basis functions by given dataset of prior information (e.g. PCA-based techniques [5] or, which is the same in this context, Karhunen-Loève expansion and its variations [3]). In majority of problems, where a model has complex structure (e.g. has to be physically consistent) the family of PCA-based techniques is used as methods which allow preserving physical consistency by incorporating two-point or multi-point correlations between elements of prior dataset [3]. Model reduction based on Karhunen-Loève expansion is also a regularization approach for inverse problems based on stochastic modelling. However, these methods still have disadvantages caused by sole using of prior information.

Key idea of this work is to reconstruct the PCA basis by introducing the information of objective function sensitivity into basis composition process.

Classic PCA basis provides best model decomposition in terms of minimizing total mean squared error of the approximation [1]:

$$\min_{\varphi_k} \mathbf{E}\Big[\int_r \big(\delta x(r)\big)^2 dr\Big] \tag{2}$$

$$\delta x(r) = x_\omega(r) - \tilde{x} = \sum_{k=1}^{\infty} A_k(\omega)\varphi_k(r) - \sum_{k=1}^{N} A_k(\omega)\varphi_k(r) = \sum_{k=N+1}^{\infty} A_k(\omega)\varphi_k(r) \tag{3}$$

where $\delta x(r)$ is the approximation error between model $x_\omega(r)$ and its projection $\tilde{x}$ onto hyperplane formed by basis function $\varphi_k(r)$. $A_k(\omega)$ are corresponding coefficients of the model decomposition.

Main proposition of this work is to include objective function sensitivity information into (2) as:

$$\min_{\varphi_k} \mathbf{E}\Big[\int_r \big(\delta x(r)\big)^2 dr + \gamma \delta F\Big] \tag{4}$$

where δF is difference between objective function of initial model and approximated model:

$$\delta F(x) = F(x) - F(\tilde{x}) = F(\tilde{x} + \delta x) - F(\tilde{x}) \tag{5}$$

As a result, the efficient numerical algorithm for calculation of update for the PCA basis was derived. The technique is easy to implement in combination with any parametrization algorithm from PCA family. It also doesn't significantly increase computational cost of the whole workflow of solving inverse problem (1). The key benefit of this technique is improved performance of classic PCA-based techniques in presence of complex model structure and improper prior info dataset.

The result was achieved by applying quantum mechanics Perturbation Theory and adjoint technique for gradient calculation [4]. In presentation, key examples of applications in History Matching problem and comparative analysis with other popular parametrization techniques (DCT, DWT, PCA family) will be provided.

References

[1] Richard Everson and Lawrence Sirovich. Karhunen–loeve procedure for gappy data. *JOSA A*, 12(8):1657–1664, 1995.

[2] Richard Wilfred Rwechungura, Mohsen Dadashpour, Jon Kleppe, et al. Advanced history matching techniques reviewed. In *SPE Middle East Oil and Gas Show and Conference*. Society of Petroleum Engineers, 2011.

[3] Pallav Sarma, Louis J Durlofsky, and Khalid Aziz. Kernel principal component analysis for efficient, differentiable parameterization of multipoint geostatistics. *Mathematical Geosciences*, 40(1):3–32, 2008.

[4] Pallav Sarma, Louis J Durlofsky, Khalid Aziz, and Wen H Chen. Efficient real-time reservoir management using adjoint-based optimal control and model updating. *Computational Geosciences*, 10(1):3–36, 2006.

[5] Sharad Yadav et al. History matching using face-recognition technique based on principal component analysis. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, 2006.

Short biography – A. Mukhin got his B.S degree from the Moscow Institute of Physics and Technology (MIPT) and is currently enrolled in M.S program in MIPT co-funded by MIPT Center for Engineering and Technology. His topic focuses on the solving complex inverse problems such as history matching.

Combining geostatistics and numerical simulations to improve estimations of pollution plumes in groundwater.

L. PANNECOUCKE
MINES ParisTech, Centre de Géosciences

Supervisor(s): C. de Fouquet (MINES ParisTech), X. Freulon (MINES ParisTech) and M. Le Coz (IRSN)

Ph.D. expected duration: Nov. 2017 - Oct. 2020

Adress: 35 rue Saint-Honoré, 77300 Fontainebleau

Email: lea.pannecoucke@mines-paristech.fr

Abstract:

Characterization of contaminated soil and groundwater around industrial plants is a major challenge for site remediation. A classical approach consists in providing an estimation of the polluted zone extent thanks to observations (data of pollutant concentration) and geostatistical tools (*e.g.* kriging). However, this estimation might turn out to be of low precision if only few data are available. Besides, flow and contaminant transport simulation is widely used to assess potential migration paths of pollutants through the subsurface. It is efficient even if information from sampling is not available, as long as input parameters are consistent with the site under study.

Thus, the approach developed in this work combines classical geostatistical tools and results of simulations of flow and contaminant transport. It aims at improving the quality of the estimation of the polluted zone extent and reducing the associated uncertainties.

The proposed method is adapted from the work of C. Roth [1]. It consists in building an *a-priori* model of the subsurface of the site under study and simulating the migration of a pollutant plume on several realisations of that model, thus obtaining several unconditionnal realisations of pollutant plume migration. Then, hundreds of these realisations are used to compute empirical covariances accounting for the spatial variability of the regionalized variable ($\mathcal{Z}$) representing the pollution under study :

$$\mathcal{C}(x,x') = \frac{1}{N}\sum_{k=1}^{N}(\mathcal{Z}_k(x) - \overline{\mathcal{Z}(x)})(\mathcal{Z}_k(x') - \overline{\mathcal{Z}(x')}) \quad (1)$$

where $\mathcal{C}(x,x')$ is the covariance value for the couple of points (x,x'), N is the number of realisations, $\mathcal{Z}_k(x)$ is the value of $\mathcal{Z}$ at x for the k-th realisation and $\overline{\mathcal{Z}(x)}$ is the average of $\mathcal{Z}(x)$ over N realisations. Hence, we are able to compute non-stationary covariances that reproduce the spatial variability of $\mathcal{Z}$ better than a model based on observations only. Finally, a kriging estimate using these non-stationnary covariances is performed. The same approach can also be applied if the spatio-temporal aspect of $\mathcal{Z}$ is considered, by computing empirical spatio-temporal covariances.

The performances of this method are assessed on a two-dimensionnal synthetic model of subsurface with a scenario of pollution due to a tritium source. The model includes an unsaturated zone of a few meters deep in which the flow and contaminant transport is simulated. Only few observations are extracted from the reference simulation (*e.g.* Figure 1) so as to evaluate the extent of the polluted zone. Then, hundreds of flow and contaminant transport simulations are run with input parameters differing from the reference simulation, in order to take into account the uncertainties on the input parameters of the modeling.

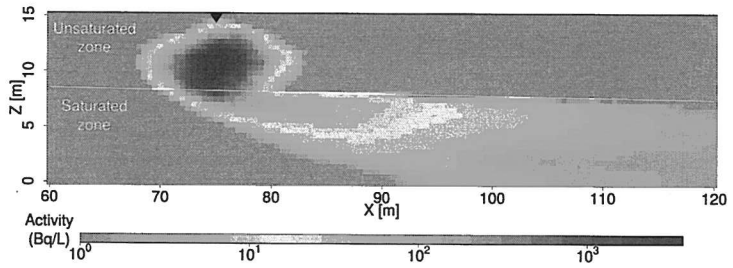


Figure 1: Example of tritium plume used as reference. The triangle highligths the pollution source point.

The extent of the reference polluted zone is then estimated using (i) a classical geostatistical method considered here as a benchmark and (ii) the above-mentioned method combining geostatistical tools and data from simulations. The results show that the estimations are improved when using data from simulations in a geostatistical modeling framework, even if few observations are available, which underlines the interest of this method.

Finally, the proposed approach could help better estimate volumes of soils to be decontaminated in the context of remediation of industrial sites. It is presented here in the context of a radiological pollution but it could be transposed to other types of pollution.

References

[1] Christopher Roth. *Contribution de la géostatistique à la résolution du problème inverse en hydrogéologie.* Thèse. École Nationale Supérieure des Mines de Paris, 1995.

Short biography – After an engineer degree at MINES ParisTech, I started a PhD thesis in applied geostatistics in november 2017. This PhD thesis is part of a project supported by Andra through the "Investments for the Future" Program. This project aims at improving the characterization of polluted soils around nuclear facilities, prior to their dismantlement. Both MINES ParisTech, IRSN and GEOPS (Paris Sud University) collaborate on this project.

Metamodelling for spatial outputs with functional PCA. Application to marine flooding.

T.V.E. PERRIN
Mines Saint-tienne (EMSE), France

Supervisor(s): O.Roustant (EMSE), J.Rohmer (BRGM), D.Moncoulon (CCR), J.Naulin (CCR), P.Tinard (CCR)

Ph.D. expected duration: Oct. 2017 - Sep. 2020

Adress: 158 Cours Fauriel, 42023 Saint-tienne

Email: elodie.perrin@emse.fr

Abstract:

This abstract is for a poster submission. Gaussian process (GP) is one of the most attractive metamodels for emulating time-consuming computer codes. Here, we focus on problems when outputs are spatial maps. Without loss of generality, we consider a spatial domain $D_z = [0,1]^2$. The computer code is viewed as a function:

$$\begin{array}{rccc} f & : & \mathcal{X} \subseteq \mathbb{R}^d & \to & \mathcal{L}_2([0,1]^2) \\ & & x & \mapsto & y_x(z) \end{array} \tag{1}$$

where x is a vector of scalar inputs and $y_x(z)$ is the ouput at the location $z \in D_z$.

A common technique [1] is to vectorize and to reduce dimensionality of the output map by principal component analysis (PCA). Then, PCA coordinates are emulated independantly by different GP models, which seems reasonable given the orhonormality of the axis. However, output dimensionality can be too high and makes intractable the covariance matrix diagonalization. For instance, marine flooding maps may have more than ten thousand pixels. Furthermore, PCA does not take into account the spatial nature of the data and their properties, such as smoothness.

In our work, $\forall x \in \mathcal{X}$, $y_x(z)$ is decomposed onto a finite basis of functions, using functional principal components analysis (FPCA) [3]. This dimension reduction method allows to keep at most the functional/spatial nature of the outputs. In the litterature, wavelets are often chosen as the basis of functions for spatial maps [2], due to their ability in revealing information at different levels and areas in the maps. Thus we have:

$$y_x(z) = \sum_{j=1}^{K} \beta_j(x)\phi_j(z), \ \forall x \in \mathcal{X}, \ \forall z \in D_z \tag{2}$$

where $j \in \{1, \ldots, K\}$, $\beta_j(x)$ are the wavelet coefficients, and $(\phi_j(z))$ the wavelet basis. This basis is orthonormal, so FPCA is equivalent to PCA on the wavelet coefficients. Obviously, in order to reduce the dimension, it is necessary to limit PCA on the most informative wavelet coefficients. We propose to order them in two different ways according to the decomposition of the energy: $||y_x||_2^2 = \int_0^1 y_x(z)^2 dz = \sum_{j=1}^{K} \beta_j(x)^2$ (equation (3)). The unselected coefficients are estimated by the mean.

$$\lambda_j = \mathbb{E}_X \left[\frac{\beta_j(X)^2}{\sum_{j=1}^{K} \beta_j(X)^2} \right] \quad \text{or} \quad \lambda_j = \frac{\mathbb{E}_X[\beta_j(X)^2]}{\sum_{j=1}^{K} \mathbb{E}_X[\beta_j(X)^2]} \tag{3}$$

The efficiencies of metamodels obtained with FPCA and PCA are compared on a dataset of 500 computer code simulations of marine flooding at Bouchôleur site, in France. Figure 1 shows their performances using 50 different learning set (chosen at random uniformly from the whole dataset) of size 50, and emulating two principal components for FPCA and PCA. The methods are assessed using the Q^2 criterion, computed at each pixel of the maps.

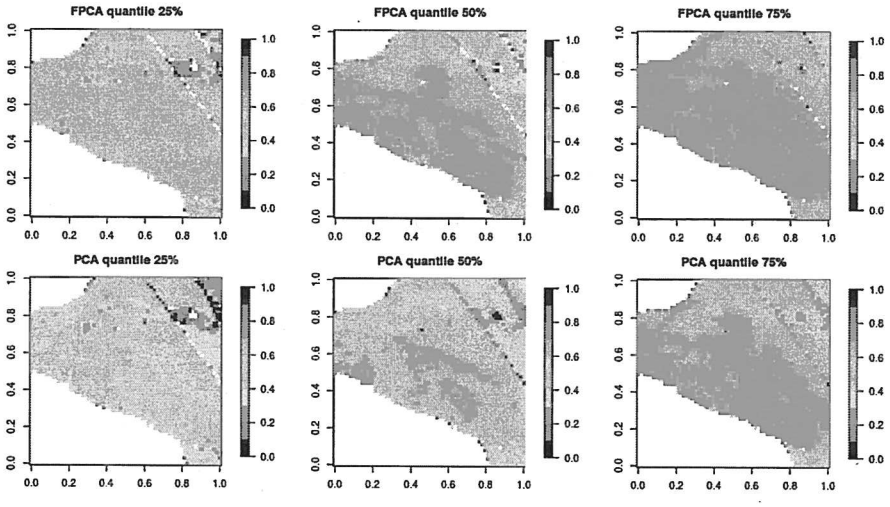


Figure 1: Q^2 maps of 25%, 50% and 75% quantiles (from left to right): metamodelling on FPCA coordinates (first line), metamodelling on PCA coordinates (second line).

It is observed that compared to the standard PCA technique, the proposed methodology leads to a better prediction accuracy (cf. Figure 1) and time computation efficiency. In our case study, applying FPCA is five times faster than PCA. On the other hand, FPCA is sensitive to the number of wavelet coefficients preselectioned, which justifies improving this aspect potentially by relying on cross-validation techniques.

References

[1] T. Chen, K. Hadinoto, W. Yan, and Y. Ma. Efficient meta-modelling of complex process simulations with time–space-dependent outputs. *Computers & chemical engineering*, 35(3):502–509, 2011.

[2] A. Marrel, B. Iooss, M. Jullien, B. Laurent, and E. Volkova. Global sensitivity analysis for models with spatialy dependent outputs. *Environmetrics*, 22(3):383–397, 2010.

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

Short biography – T.V.E. Perrin graduated the MapI^3 master in applied mathematics from Paul Sabatier University, at Toulouse, France. Now, she is doing a PhD in applied mathematics at EMSE, France, on *Propagation of uncertainties and calibration of numerical simulations to estimate the cost linked to marine flooding.* The thesis is supervised by EMSE, BRGM, and CCR.

Robust Uncertainty Quantification of a Risk Measurement from a Computer Code

J. STENGER
Université Toulouse III - Paul Sabatier

Supervisor(s): Prof. F. Gamboa (IMT), Dr. M. Keller (EDF) and Dr. B. Iooss (EDF)

Ph.D. expected duration: Oct. 2017 - Sep. 2020

Adress: EDF R&D, 6 quai Chatier, 78400 Chatou

Email: jerome.stenger@edf.fr

Abstract:

Uncertainty quantification methods address problems related to with real world variability. Generally, an engineering system is represented by a numerical function $Y = G(X)$, whose inputs $X \in \mathbb{R}^p$ are uncertain and modeled by random variables. The variable of interest is the scalar output of the computer code, but the statistician rather work with some quantities of interest, for example, a quantile, a probability of failure, or any measures of risk. Uncertainty quantification aims to characterize how the variability of a system and its model affect the quantity of interest [1].

We propose to gain robustness on the quantification of this measure of risk. Usually input values are simulated from an associated joint probability distribution. This distribution is often chosen in a parametric family, and its parameters are estimated using a sample and/or the opinion of an expert. However the difference between the probabilistic model and the reality induces uncertainty. The uncertainty on the input distributions is propagated to the quantity of interest, as a consequence, different choices of input distributions will lead to different values of the risk measures.

To consider this uncertainty, we propose to evaluate the maximum risk measure over a class of distributions. Different classes are suggested in the literature mainly discussed in the work of Berger and Hartigan in the context of Robust Bayesian Analysis (see [10]). They consider for example the generalized moment set [4] or the ε-contamination set [11]. The generalized moment set has some really nice properties studied by Winkler [13] based on the well known Choquet theory [3]. An extension of Winkler's work has been more recently published by Owhadi and al. [9] under the name of Optimal Uncertainty Quantification (OUQ). In our work, we will focus on classes of measures specified by classical moment constraints. This is a particular case of the framework introduced by [9] justified by our industrial context, mainly related to nuclear safety issues [12]. Indeed, in practice the estimation of the input distributions, built with the help of the expert, often relies only on the knowledge of the mean or the variance of the input variables.

The solution of our optimization problem is numerically computed thanks to the OUQ reduction theorem ([9], [13]). This theorem states that the maximum of the risk measure is located on the extreme points of the distribution set. In the context of the moment class, it corresponds to a product of discrete finite measures. To be more specific it holds that when N pieces of information are available on the moments of a measure μ, it is enough to pretend that the measure is supported on at most $N + 1$ points.

One of the main issues is the computational complexity of the optimization of the risk measure over the given class of distribution. In the moment context, Semi-Definite-Programming [6] has been already already explored by Betrò [2] and Lasserre [7], but the deterministic solver rapidly reaches its limitation as the dimension of the problem increases. One can also find in the literature

a Python toolbox developed by McKerns and al. [8] called Mystic framework that fully integrates the OUQ framework. However, it was built as a generic tool for generalized moment problems and the enforcement of the moment constraints is not optimal. By restricting the work to classical moment sets, we propose an original and practical approach based on the theory of canonical moments [5]. Canonical moments of a measure can be seen as the relative position of its moment sequence in the moment space. It is inherent to the measure and therefore present many interesting properties. Our algorithm shows very good performances and great adaptability to any constraints order. The dimension is subject to the curse of dimension but can be perform up to dimension 10 for a reasonable cost.

References

[1] Michal Baudin, Anne Dutfoy, Bertrand Iooss, and Anne-Laure Popelin. Open TURNS: An industrial software for uncertainty quantification in simulation. In D. Higdon R. Ghanem and H. Owhadi, editors, *Handbook of uncertainty quantification*. Springer, 2017.

[2] Bruno Betrò. *Robust Bayesian Analysis*, chapter 15, Methods for Global Prior Robustness under Generalized Moment Conditions. Lecture Notes in Statistics. Springer-Verlag, New York, 2000.

[3] Gustave Choquet, Jerrold Marsden, and Stephen Gelbart. Lectures on analysis / Gustave Choquet. *SERBIULA (sistema Librum 2.0)*, July 2018.

[4] Lorraine DeRoberts and J. A. Hartigan. Bayesian Inference Using Intervals of Measures. *The Annals of Statistics*, 9(2):235–244, March 1981.

[5] Holger Dette and William J. Studden. *The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis*. Wiley-Blackwell, New York, September 1997.

[6] Didier Henrion, Jean-Bernard Lasserre, and Johan Lfberg. GloptiPoly 3: moments, optimization and semidefinite programming. *Optimization Methods and Software*, 24(4-5):761–779, October 2009.

[7] Jean-Bernard Lasserre. *Moments, positive polynomials and their applications*. Number v. 1 in Imperial College Press optimization series. Imperial College Press ; Distributed by World Scientific Publishing Co, London : Signapore ; Hackensack, NJ, 2010. OCLC: ocn503631126.

[8] M. McKerns, H. Owhadi, C. Scovel, T. J. Sullivan, and M. Ortiz. The optimal uncertainty algorithm in the mystic framework. *CoRR*, abs/1202.1055, 2012.

[9] Houman Owhadi, Clint Scovel, Timothy John Sullivan, Mike McKerns, and Michael Ortiz. Optimal Uncertainty Quantification. *SIAM Review*, 55(2):271–345, January 2013. arXiv: 1009.0679.

[10] Fabrizio Ruggeri, David Rios Insua, and Jacinto Martin. Robust Bayesian Analysis. In D. K. Dey and C. R. Rao, editors, *Handbook of Statistics*, volume 25 of *Bayesian Thinking*, pages 623–667. Elsevier, January 2005.

[11] S. Sivaganesan and James O. Berger. Ranges of Posterior Measures for Priors with Unimodal Contaminations. *The Annals of Statistics*, 17(2):868–889, June 1989.

[12] G.B. Wallis. Uncertainties and probabilities in nuclear reactor regulation. *Nuclear Engineering and Design*, 237:1586–1592, 2004.

[13] Gerhard Winkler. Extreme Points of Moment Sets. *Math. Oper. Res.*, 13(4):581–587, November 1988.

Short biography – The industrial needs of statistical robustness are required by the IRSN in the context of nuclear safety. This PhD originates from the close colaboration of EDF R&D and the University of Toulouse III - Paul Sabatier. Key words: *robustness, uncertainty quantification, optimization, canonical moments.*

Optimisation of multi-year planning strategies to better integrate renewable energies and new electricity uses in the distribution grid

B. Tebbal Barracosa
L2S, CentraleSupélec, University of Paris-Saclay, EDF R&D

Supervisor(s): Prof. Emmanuel Vazquez (L2S, CentraleSupélec, University of Paris-Saclay), Prof. Julien Bect (L2S, CentraleSupélec, University of Paris-Saclay), Dr. Héloïse Baraffe (EDF R&D) and Dr. Juliette Morin (EDF R&D)

Ph.D. expected duration: Nov. 2018 - Oct. 2021

Address:
EDF Lab Paris-Saclay, 7 Boulevard Gaspard Monge
Bât OPALE – 1^{er} étage – Bureau O1B.24
91120 Palaiseau

Email: bruno.tebbal-barracosa@edf.fr

Abstract:

The planning of electricity distribution grids is expected to evolve quickly to not only incorporate an increasing number of electricity production facilities based on Renewable Energies (RE) but also to anticipate the introduction of new electricity uses such as electric vehicles and energy storage equipment.

This new reality can cause the need for work on the grid that is costly and long to implement. Different research centres, including EDF R&D, assign resources to the development of solutions capable of easing the integration of REs and new electricity uses into the distribution grid: advanced voltage regulation, temporary renewable production curtailment ...

EDF R&D and CentraleSupélec/L2S cooperate since 2012 to develop a decision support tool for the multi-year planning of distribution grids with the introduction of renewable energies [1]. Using this tool and for a given network, different planning strategies representing different approaches can be studied, while taking into account uncertainties regarding the deployment of renewable energies. EDF R&D is interested in using this tool to identify the best planning strategies to apply to a given family of networks (rural networks, urban ...).

Until now the research into the optimisation of planning strategies has consisted in minimising the expected cost of a bi-variable strategy through the use of Bayesian optimisation algorithms such as IAGO [4], TTPS and PTS [3]. The tested algorithms showed interesting performances in a simple case but do not seem however entirely adapted to the future needs in terms of optimisation:

- Expected cost may not be the best criterion for measuring the effectiveness of a strategy. Statistical indicators that are more relevant but also more difficult to estimate should be considered;
- The optimisation of a single objective, such as the average cost, is not satisfactory if one considers the complexity of the problem on the Distribution System Operator (DSO) perspective. Multi-objective and/or constrained formulations must, therefore, be studied [2];
- A DSO eventually prefers to know the near optimal areas of the solution space, satisfying a tolerance on objectives, rather than the exact solution. This way of considering optimisation does not seem to have been considered in the literature.

In order to develop the existing tool several steps were identified: model new multi-variable planning strategies, including new technical solutions; define criteria to measure the effectiveness of the planning strategies; propose different possible formulations of the optimisation problem, including several objectives and/or constraints potentially contradictory and/or difficult to estimate; develop optimisation algorithms adapted to the peculiarities of the problem; demonstrate the interest of algorithms in case studies, including different REs penetration rates, planning strategies and distribution network families.

In this presentation, the overall problem and the initial exploratory work regarding criterion to measure the effectiveness of planning strategies shall be presented.

References

[1] Héloïse Dutrieux. *Méthodes pour la planification pluriannuelle des réseaux de distribution. Application à l'analyse technico-économique des solutions d intégration des énergies renouvelables intermittentes.* PhD thesis, Ecole centrale de Lille, 2015.

[2] Paul Feliot, Julien Bect, and Emmanuel Vazquez. A bayesian approach to constrained single- and multi-objective optimization. *Journal of Global Optimization*, 67(1-2):97–133, 2017.

[3] Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

[4] Julien Villemonteix, Emmanuel Vazquez, and Eric Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

Short biography – With a background in Industrial Engineering and Management (University of Lisbon), and a Master Degree in Renewable Energies (Ecole Polytechnique), Bruno began his PhD in November 2018 with CentraleSupélec and EDF R&D.
This PhD thesis is funded by EDF R&D, under the scope of a CIFRE Agreement, and aims at further developing EDF R&D tools for the planning of the Electricity Distribution Grids.

A rigorous framework to describe margins

ADRIEN TOUBOUL
CERMICS - Université Paris-Est

Supervisor(s): Pr. Bernard Lapeyre (CERMICS), Dr. Julien Reygner (CERMICS), Dr Mouadh Yagoubi (IRT SystemX), Mr. Fabien Mangeant (Renault)

Ph.D. expected duration: Nov. 2017 - Nov. 2020

Adress: IRT SystemX - 8 Avenue de la Vauve, 91120 Palaiseau

Email: adrien.touboul@irt-systemx.fr

Abstract: The concept of margin is widely used in engineering fields, when the topic of design under uncertainty needs to be addressed. This notion has first been shaped by the intuition, but more rigorous definitions have been proposed in some engineering fields, as the practices were understood more precisely. Nevertheless, it appears that there is no *general* definition that would describe formally a margin independently from the field [2]. The goal of this work is to provide such a framework, encompassing a wide variety of current practices.

The margins that are investigated are those that can be described as *an amount of something included so as to be sure of success or safety.* Considering this definition, the margins in our scope are always taken to cover the consequences of some uncertainties. Our thesis is that it is always possible to measure this *amount of something included*, as a distance to a reference. However, this reference is not always a critical point to avoid or a point of performance to aim at and the distance does not always include all the variables of the problem.

The source of the uncertainty (lack of knowledge, truly aleatory phenomenon, unkown future design choices, unreliable partners...) and the consequences (limiting the design, improper prediction, safety concerns...) at stake in margins are numerous and diverse. In order to model it, it is assumed that it is possible to determine if the system is in an acceptable state by looking at some (random) variables of interest describing the system. More precisely, a state is acceptable if and only if the random variables of interest belong to an acceptance set $\mathcal{A}$. This acceptance set can be defined thanks to a risk threshold and a risk measure - i.e a function that maps random variables to a real value, interpreted as a "risk"-, similarly to the monetary risk measures introduced in [1].

We show that our framework gives a relevant interpretation of some well documented indicators from multiple engineering fields, the capability process C_p in statistical quality [7], the safety coefficients γ in probabilistic civil engineering [5], the gain GM and phase margin PM in control [6], among others. Some specific margins defined in previous frameworks [8], [4], [3] can also be defined in our formal system.

One of the possible applications of such a framework is to permit communication and exchange of margins between engineering disciplines, in the context of the design of a complex system. Another use could be an easier recording of the reason why a margin is imposed and an easier monitoring of its future evolution. The study of margin calibration, i.e focusing on how the minimum margin values are chosen, could be facilitated by the proposed margin definition. Last but not least, we hope that it will help to formalize some problems known as global margin allocation and margin accumulation, that would be solved by UQ techniques.

References

[1] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[2] Claudia Eckert and Ola Isaksson. Safety margins and design margins: A differentiation between interconnected concepts. *Procedia CIRP*, 60:267 – 272, 2017. Complex Systems Engineering and Development Proceedings of the 27th CIRP Design Conference Cranfield University, UK 10th – 12th May 2017.

[3] Marin D. Guenov, Xin Chen, Arturo Molina-Cristóbal, Atif Riaz, Albert S. J. van Heerden, and Mattia Padulo. Margin allocation and tradeoff in complex systems design and optimization. *AIAA Journal*, 56(7):2887–2902, December 2018.

[4] Jon C. Helton. Quantification of margins and uncertainties: Conceptual and computational basis. *Reliability Engineering & System Safety*, 96(9):976 – 1013, 2011. Quantification of Margins and Uncertainties.

[5] Maurice Lemaire. *Structural reliability*. John Wiley & Sons, 2013.

[6] William S Levine. *The Control Systems Handbook: Control System Fundamentals*. CRC press, second edtition edition, 2010.

[7] Douglas C Montgomery. *Introduction to statistical quality control*. John Wiley & Sons, s edition, 2007.

[8] Daniel P Thunnissen and Glenn T Tsuyuki. Margin determination in the design and development of a thermal control system. 2004.

Short biography – Adrien Touboul completed a Master's degree in applied mathematics at University Pierre et Marie Curie as well as an Engineer degree at École des Ponts ParisTech in 2017. He is now pursuing a PhD at the research institute IRT SystemX in a team gathering multiple industrial players, coming from PSA group, Renault and Airbus, among others. The academic support for the PhD is provided by the research center in applied mathematics CERMICS.

Emulating the response PDF of stochastic simulators using sparse generalized lambda models

X. Zhu
ETH Zürich

Supervisor(s): Prof. Dr. B. Sudret (ETH Zürich)

Ph.D. expected duration: Oct. 2017 - Sep. 2021

Adress: Chair of Risk, Safety and Uncertainty Quantification, Stefano-Franscini-Platz 5, 8093, Zürich, Switzerland

Email: zhu@ibk.baug.ethz.ch

Abstract:

With increasing demands on the functionality of structures, more and more complex interdependent infrastructures and networks are developed in engineering. Design and maintenance of such systems require advanced computational models (a.k.a. simulators) to assess the reliability, control the risk and optimize the behaviour of the systems. Classically, numerical models are *deterministic*, meaning that repeated model evaluations with the same input parameters produce exactly the same output quantity of interest (QoI). In contrast, *stochastic simulators* provide different results when run twice with the same input values. In other words, the QoI of a stochastic simulator is a random variable for a given vector of input parameters. The case study that fosters this research work is the structural design of wind turbines: the input of the simulator is among others, a three-dimensional wind field, which is macroscopically defined only with a few parameters. A single realization of those parameters leads to different realizations of the wind field, and thus to different structural performance.

In the context of optimization or uncertainty quantification, surrogate models are often used to alleviate the computational burden. Deterministic surrogate methods have been successfully developed in the past two decades, yet they cannot be directly applied to emulate stochastic simulators due to the random nature of the latter. To build stochastic emulators, two categories of methods can be found in the literature. The first one focuses on estimating some statistical scalar quantities, *e.g.* mean and variance [2]. The second category aims to estimate the response distribution function but requires a large size of the data set, especially many replications of the runs of the simulator to capture the intrinsic stochasticity of the response [3]. In this work, we introduce a novel approach that does not require replications to build a sparse surrogate model that predicts the response probability density function (PDF) for any input parameter set.

For given input parameters $\boldsymbol{X} = \boldsymbol{x} \in \mathbb{R}^M$, we choose to approximate the PDF of the QoI $Y(\boldsymbol{x})$ using the four-parameter generalized lambda distribution (GLD) [1]. Following this setting, the distribution parameters $\boldsymbol{\lambda}$ are functions of the input parameters, *i.e.* $\boldsymbol{\lambda}(\boldsymbol{X})$. Under certain conditions, these functions can be represented by polynomial chaos (PC) expansions [4], and the coefficients associated with the PC basis functions are the model parameters to be estimated from data. In summary, the model is expressed as follows:

$$Y(\boldsymbol{X}) \sim GLD(\lambda_1(\boldsymbol{X}), \lambda_2(\boldsymbol{X}), \lambda_3(\boldsymbol{X}), \lambda_4(\boldsymbol{X})) \tag{1}$$

$$\lambda_s(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} a_{s,\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \quad s = 1, 2, 3, 4 \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_M)$ denotes the multi-index defining the PC basis function $\psi_{\boldsymbol{\alpha}}(\boldsymbol{X})$ and $a_{s,\boldsymbol{\alpha}}$ is the associated coefficient with respect to $\lambda_s(\boldsymbol{X})$.

When fitting the GLD model to data, truncated series expansions using a finite set of multivariate orthogonal polynomials $\{\psi_\alpha, \alpha \in \mathcal{A}_s\}$ is used for each $\lambda_s(\boldsymbol{X})$. If prior knowledge is available to fix $\mathcal{A}_s$, we propose the maximum likelihood estimation to estimate the model parameters $\boldsymbol{a}$ without the need for replications. However, due to the complexity of the GLD formulation, maximizing the likelihood can be time consuming with large data sets. Here, we derive analytically the gradient and Hessian matrix of the likelihood function with respect to $\boldsymbol{a}$ to efficiently apply derivative-based optimization algorithms. In the case of unknown $\mathcal{A}_s$, the classical "full" PC approximation cannot be applied due to the so-called *curse of dimensionality*, in the sense that the basis size increases exponentially with increasing input dimension or polynomial degree [4]. To overcome the difficulty, we propose a stepwise algorithm to adaptively construct a sparse PC approximation for each $\lambda_s(\boldsymbol{X})$. The method mainly consists of three steps: estimation of the conditional mean and variance, forward selection and backward elimination.

The performance of the proposed method is illustrated in various analytical examples. Further applications to wind turbine simulations is currently in progress.

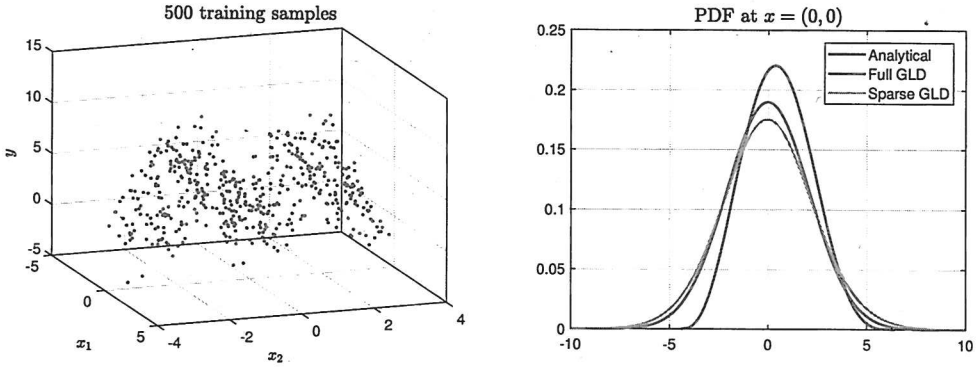


Figure 1: On the left: training data. On the right: comparison between the GLD model built with full PC approximations of $\boldsymbol{\lambda}(\boldsymbol{X})$ (denoted by full GLD) and the sparse GLD model. This toy example has normal distribution as the analytical solution with its mean and variance depending on the two-dimensional input parameters.

References

[1] M. Freimer, G. Kollia, G. S. Mudholkar, and C. T. Lin. A study of the generalized Tukey lambda family. *Comm. Stat. Theor. Meth.*, 17:3547–3567, 1988.

[2] A. Marrel, B. Iooss, S. Da Veiga, and M. Ribatet. Global sensitivity analysis of stochastic computer models with joint metamodels. *Stat. Comput.*, 22:833–847, 2012.

[3] V. Moutoussamy, S. Nanty, and B. Pauwels. Emulators for stochastic simulation codes. *ESAIM: Math. Model. Num. Anal.*, 48:116–155, 2015.

[4] B. Sudret. *Polynomial chaos expansions and stochastic finite element methods*, chapter 6, pages 265–300. Risk and Reliability in Geotechnical Engineering. Taylor and Francis, 2015.

Short biography – Xujia Zhu received his engineer degree from Ecole Polytechnique (France) in 2015. He also holds a MSc in computational mechanics from the Technical University of Munich. Since October 2017, he is a Ph.D. student at the Chair of Risk, Safety and Uncertainty Quantification with the thesis entitled "*Surrogate modelling for stochastic simulators using statistical approaches*" funded by the Swiss National Science Foundation.