Oral sessions
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


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).
Address: 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 $Y = M(X_1, ..., X_d)$ where the scalar output $Y$ depends on a $d$-dimensional real valued random variable $X = (X_1, ..., X_d)$ through a deterministic function $M$. We assume that for every $I \subset \{1, ..., d\}$ a strict subset, the pair $(X_I, Y)$ is absolutely continuous. The idea of Borgonovo’s GSA approach is to measure how fixing $X_I$ at a value $x_I$ modifies the entire distribution of the output $Y$. This modification is quantified by the shift $s(x_I)$ defined from the $L_1$-norm between the output probability density function (PDF) $f_Y$ and the conditional output PDF $f_{Y|X_I=x_I}$:

$$s(x_I) = \frac{1}{2} \left\| f_Y - f_{Y|X_I=x_I} \right\|_{L_1(\mathbb{R})} = \frac{1}{2} \int \left| f_Y(y) - f_{Y|X_I=x_I}(y) \right| dy.$$  \hspace{1cm} (1)

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

$$\delta_I := E [s(x_I)] .$$ \hspace{1cm} (2)

Estimating Borgonovo’s indices is a challenging task because of the unknown unconditional and conditional PDFs $f_Y$ and $f_{Y|X_I=x_I}$ 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 $\delta_i$ 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:

$$\delta_I = \frac{1}{2} \left\| f_{X_I,Y} - f_{X_I} f_Y \right\|_{L_1(\mathbb{R}^2)} .$$ \hspace{1cm} (3)

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


Address: 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


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).
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)


Address: 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 \(x \geq x'\)).

![Diagram of GP regression models](image.png)

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...
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 Pouchardier, 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 \( f(x) \) where \( x \in X \subset \mathbb{R}^d \) is a \( d \)-dimensional vector of CAD parameters. It is common that \( d \) is large, \( d \gg 50 \). 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 \( d' \ll d \) variables.

In this work, we exploit the fact that the computation time of a shape \( \Omega_x \) is negligible in comparison with the evaluation time of \( f(x) \). Most often, the set of all CAD generated shapes, \( \Omega := \{\Omega_x, x \in X\} \) can be approximated in a \( \delta \ll d \)-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 \( \phi \) 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 \( \phi \), we have found that a few \( (\delta) \) eigenshapes allow to accurately describe the sample of CAD shapes through their principal components, \( a \) in the eigenbasis \( (v^1, \ldots, v^D) \).

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

\[
Y(a) = \mu + Y^a(a_{1:\delta}) + Y^\pi(a_{\delta+1:D}) + \varepsilon.
\]

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

Address: 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 \( f: \mathbb{X} \times \mathbb{V} \to \mathbb{R} \) 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 \( \mathbb{V} \). We assume that the probability distribution of the functional input is only known through a finite set of realizations \( \Xi = \{ v_1, \ldots, v_n \} \) and each evaluation of \( f \) involves a time consuming computation. Let \( X_n = \{ x_1, \ldots, x_n \} \) and \( V_n = \{ v^1, \ldots, v^n \} \) denote the initial design points. The simulator responses at these design points are denoted by \( Z = \{ f(x_1, v^1), \ldots, f(x_n, v^n) \} \).

The objective of this study is to estimate the set

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

where \( c \in \mathbb{R} \) and \( \mathbb{E}_\mathbb{V} \) is the expectation with respect to \( \mathbb{V} \). The estimation of \( \Gamma^* \) 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 \( (\mathbb{V} \in \mathcal{V} \rightarrow \mathbb{U} \in \mathbb{R}^{m_u}) \), a GP model \( Z_{(x,u)} \) is built in the joint space of control and uncertain variables \( (x, u) \). Then an averaged GP over the random variables is derived [2], \( Y_{(x)} = \mathbb{E}_U[Z_{(x,u)}] \). Therefore the ‘induced’ Gaussian process approximates the expected response involved in inversion problem (1). Driven by the Gaussian process conditioned on the \( n \) observations \( Y_{(x)}^{(n)} \), the proposed infill strategy consists of two steps: first we choose \( x^{n+1} \) by minimizing the Vorob’ev deviation [1]. Secondly we choose the uncertain point \( u^{n+1} \) that minimizes the variance of the process \( Y_{(x)}^{(n)} \) evaluated at the point \( x^{n+1} \). By iterating this procedure, we expect a rapid estimation of the set \( \Gamma^* \) with a limited number of calls to the simulator \( f \).
Sequential polynomial chaos expansions for Bayesian posterior computations with expensive computational forward models

P.-R. Wagner
ETH Zurich

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


Address: 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 $X \sim \pi(x)$ following the observation of measurements $y$ through the well known Bayes’ theorem:

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

(1)

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

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 $\mathcal{L}(x;y)$. 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 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.
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

Address: 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 \( \Omega \) 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 \( P(Y|X=x) \). 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 \( P(Y|X=x) \) may vary with respect to \( x \) (heteroscedasticity) and that \( P(Y|X=x) \) does not belong to any specific parametric class.

The conditional quantile function of order \( \tau \) is defined as \( q_\tau(x) = \inf \{ q : F_q(X=x) \geq \tau \} \), \( \tau \in (0,1) \), where \( F_q(X=x) \) is the cumulative distribution function of \( P(Y|X=x) \), 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 \( T \). 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 \( T_1 \) 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
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.
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:


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 priori knowledge 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.
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.
Abstract

Active observation of Level Lines of Spatial Fields

We present a design criterion for the active acquisition of level sets of a spatial field. Current adaptive design algorithms for the estimation of the probability that a field exceeds a certain value greedily select the observations that optimize uncertainty of the estimator of exceedance probability, relying on a prior Gaussian Process model for the field. Although these design methods do induce a clustering of the design points in the vicinity of the target level set of the spatial field, they are not designed for the situations where the main interest is to spatially locate the level line. This is intuitively understood by observing that uncertainty about the locus of the target level curve strongly depends on the internal correlation of the predicted field, while current design criteria only consider the marginal distributions, and thus attached uncertainties, at each point of the field’s domain.
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.
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.
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,

$$H_q(D) = \frac{1}{1-q} \ln \int_{[0,1]^d} f(x)^q dx$$

measures the difference between $f$ and the uniform density function in so far as, one always has $H_q(D) \geq 0$ and the maximum value of $H_d(D)$, 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].
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:

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].


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
Abstract

A few elements of numerical analysis for PDEs with random coefficients of lognormal type

In this talk, we will address some basic issues appearing in the theoretical analysis of numerical methods for PDEs with random coefficients of lognormal type. To begin with, such problems will be motivated by applications to the study of subsurface flow with uncertainty. We will then give some results concerning the spatial regularity of solutions of such problems, which of course impacts the error committed in spatial discretization. We will complete these results with integrability properties to deal with unboundedness of these solutions and then give error bounds for the finite element approximations in adequate norms. Then we will see how these two results enable, among other things, to provide a bound for the computational cost of the multi-level Monte Carlo algorithm. Finally we will discuss the question of the dimensionality, which is crucial for numerical methods such as stochastic collocation. We will consider the approximation of the random coefficient through a Karhunen-Loève expansion, and provide bounds of the resulting error on the solution by highlighting the interest of the notion of weak error.
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
LE RICHE Rodolphe  
Ecole des Mines de Saint-Etienne  

Abstract

Restricting ambitions in global multi-objective optimization of costly functions

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 describe two ways, both involving Gaussian processes, to restrict ambitions in order to recover problems that can be solved. First, in multi-objective optimization, we discuss targeting the center of the Pareto front: what is a Pareto front center? How to detect it? What to do once it is found? Second, we provide elements of sensitivity analysis specific to optimization to freeze some of the variables: what variables count in optimization? How to fix the other variables? This talk summarizes joined works with David Gaudrie, Adrien Spagnol, Sébastien Da Veiga and Victor Picheny.
Abstract

Statistical 3D analysis of foam bubbles in porous 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.
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.
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.
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 $\omega$ in the sample space $\Omega$. 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 $\lambda_i$ and $\phi_i$ be respectively its eigenvalues and eigenvectors. Then the KL expansion [4] reads as follow:

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

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, $\lambda_i$ and $\phi_i$ 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 $\hat{H}(x^*, \omega) = \sum_{i=1}^{M} \sqrt{\lambda_i} \xi_i(\omega) \phi_i(x)$.

- $\xi_i(x^*)$ is unknown, the eigenvectors are only computed for the DoE set. To overcome this limitation, a $\phi_i$ 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 $\phi_i$, $i \in \{1 \ldots M\}$ [4]

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

Address:  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 outputs, 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 perturbation, 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 perturbation and represent strong patterns. They are selected to form a stable rule ensemble model. Our proposed algorithm works as follows:

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.
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

Address: 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 measure 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.
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)
Address: 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 at 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 \( \omega \) 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, \( \alpha \)-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

\[
u^* \in \arg \min_{u \in U} J(u) \text{ where } J(u) = \mathbb{E}(j(u,W)) \]

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.
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

Address: 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].
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)


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 $\xi$, and on the marine environment conditions, represented by the random process $Z$. This problem can be mathematically stated as

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

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 ;
- $T$ the tension of the mooring line ;
- $\beta$ the constraint on the pitch of the floater (rotation around the vertical axis) that must be less than $6^\circ$ ;
- $Z = (Z(t))_{t \in [0, T]}$ represents the sea elevation and is a locally stationary process that is piecewise stationary over intervals of duration $\Delta T : Z(t) = \sum_i \mathbb{1}_{[T_i, T_{i+1}]}(t) \zeta_i(t - T_i)$ where $T_i = i \Delta T_i$, $(\zeta_i(t))_{t \in [0, \Delta T]}$ 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
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. Ceramiczynski (Renault)


Address: 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 continuously (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), music (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 fixed 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 $T$, 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 $T$, but only on a random compact subset of $T$. So, an observation of a scene can be represented as a random vector of functions:

$$Z = \left(Z^{(1)}, \ldots, Z^{(P)} \right), \quad \forall i \in \llbracket 1, P \rrbracket, Z^{(i)} : T^{(i)} \subset T \longrightarrow \mathbb{R}^D.$$ 

Moreover, $Z^{(i)}$ is assumed to be in $L^2(T^{(i)})$ for all $i \in \llbracket 1, P \rrbracket$. So, realizations of $Z$ 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.
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
Address: 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 (\cite{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 Helmholtz equation:

\[ H(x, \xi) u = \Delta u + \frac{\omega^2}{c(x, \xi)^2} u = s(\omega) \]  \hspace{1cm} (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 parameters \( \xi \).

Linear operator theory ensures that the eigenvectors \( (\Psi_k)_{k \in K} \) of \( H \) form a basis of the space of square 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 \( (H(x, \xi)) \) and depends on the stochastic parameters \( \xi \). 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 \( (\lambda_k(\omega, \xi))_{k \in K} \) and \( (\Psi_k(x, \omega, \xi))_{k \in K} \) are computed, they can be used to generate signals for a given source at a distance \( R \):

\[ u(\omega, R, \xi) = \left[ i \sum_{k \in K} H_0^{(1)}(\lambda_k(\omega, \xi) R) \Psi_k^2(0, \omega, \xi) \right] s(\omega) \]  \hspace{1cm} (2)
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 $X = X_1 \times \cdots \times X_d$ 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 $v_\alpha$, seen as a vector-valued map, is associated to each node $\alpha$ of the dimension partition tree $T$, and this set of tensors totally parameterizes the approximation.

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 $\phi_\nu : X_\nu \rightarrow \mathbb{R}^{n_\nu}, \nu \in \{1, 2, 3, 4\}$ 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
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)


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


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 GUE-NEAU (CEA) and Thierry KLEIN (ENAC)


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 $\{x_l, z_l\}_{l=1,...,N}$, we aim at building a regularized $\mathbb{R}^p$-valued function $f = (f^1 \ldots f^p)$ such that

$$\sum_{i=1}^p \lambda^i(x_l)f^i(x_l) = z_l, \quad l = 1, \ldots, N, \quad (1)$$

with given functions $\lambda^i$. 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

$$\min_{P} K(P, P_0)$$

$$P : \int_U ydP(y) = z, \quad z = (z_1 \ldots z_N)^T \in \mathbb{R}^N. \quad (2)$$

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

$$\nu_n = \frac{1}{n} \sum_{i=1}^n X_i \delta_t, \quad (3)$$

where $X_i$ are random amplitudes and $\delta_t$ 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
Sparse polynomial chaos expansions: Benchmark of compressive sensing solvers and experimental design techniques

NORA LÜTHEN
ETH Zurich, Switzerland

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):
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)


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

A. Mukhin

Moscow Institute of Physics and Technology

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 various optimization techniques. Objective function in such optimization problems is typically written in form of squared misfit between observed data and model simulation results:

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

where $d_{i}^{\text{obs}}$ - observed data, $x$ - model, $f(x_i)$ - function of model, $\sigma_i$ - 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]:
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

Address: 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 unconditional 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 (Z) representing the pollution under study:

\[
C(x, x') = \frac{1}{N} \sum_{k=1}^{N} (Z_k(x) - \overline{Z(x)})(Z_k(x') - \overline{Z(x')})
\]

(1)

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

The performances of this method are assessed on a two-dimensional 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.
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:

$$f : \mathcal{X} \subseteq \mathbb{R}^d \rightarrow L_2([0, 1]^2)$$

where $x$ is a vector of scalar inputs and $y_x(z)$ is the output 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 independently by different GP models, which seems reasonable given the orthonormality 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 literature, 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:

$$y_x(z) = \sum_{j=1}^{K} \beta_{j}(x) \phi_{j}(z), \ \forall x \in \mathcal{X}, \ \forall z \in D_z$$

(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:

$$\|y_x\|^2 = \int_{0}^{1} y_x(z)^2 dz = \sum_{j=1}^{K} \beta_{j}(x)^2$$

(equation (3)). The unselected coefficients are estimated by the mean.

$$\lambda_{j} = \mathbb{E}[\frac{\beta_{j}(X)^2}{\sum_{j=1}^{K} \beta_{j}(X)^2}] \ \text{or} \ \lambda_{j} = \frac{\mathbb{E}[\beta_{j}(X)^2]}{\sum_{j=1}^{K} \mathbb{E}[\beta_{j}(X)^2]}$$

(3)
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)
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 $\varepsilon$-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 $\mu$, 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...
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)


Address: EDF Lab Paris-Saclay, 7 Boulevard Gaspard Monge
Bât OPALE – 1er 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.
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

Address: 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, unknown 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 $\gamma$ 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

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

X. Zhu  
ETH Zürich

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

Address: 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 $X = x \in \mathbb{R}^M$, we choose to approximate the PDF of the QoI $Y(x)$ using the four-parameter generalized lambda distribution (GLD) [1]. Following this setting, the distribution parameters $\lambda$ are functions of the input parameters, i.e. $\lambda(X)$. 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:

$$Y(X) \sim GLD(\lambda_1(X), \lambda_2(X), \lambda_3(X), \lambda_4(X))$$

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

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_M)$ denotes the multi-index defining the PC basis function $\psi_\alpha(X)$ and $a_{s, \alpha}$ is the associated coefficient with respect to $\lambda_s(X)$. 
When fitting the GLD model to data, truncated series expansions using a finite set of multivariate orthogonal polynomials \( \{ \psi_{\alpha}, \alpha \in A \} \) is used for each \( \lambda_s(\mathbf{X}) \). If prior knowledge is available to fix \( A_s \), we propose the maximum likelihood estimation to estimate the model parameters \( a \) 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 \( a \) to efficiently apply derivative-based optimization algorithms. In the case of unknown \( A_s \), 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 \( \lambda_s(\mathbf{X}) \). 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 \( \lambda(\mathbf{X}) \) (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.](image)

References


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.