MascotNum 2024 PhD Students Day

Book of Abstracts

Talks & Posters

PhD Students Talks

PhD Students Posters

Adaptive importance sampling of stochastic processes with graph-based mean hitting times

Guillaume Chennetier[†,1,2], Hassane Chraibi[1], Anne Dutfoy[1], Josselin Garnier[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Apr. 2021 – May. 2024

[1] EDF Lab Paris-Saclay, Boulevard Gaspard Monge, 91120 Palaiseau, France
{guillaume.chennetier,hassane.chraibi,anne.dutfoy}@edf.fr

[2] CMAP, Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France
{guillaume.chennetier,josselin.garnier}@polytechnique.edu

Abstract

Importance sampling is one of the most famous and perhaps the most direct method of variance reduction in simulation-based inference when estimating quantities of the form $\mathbb{E}_{X\sim\mathbf{p}}[\varphi(X)]$.

$$\overline{\varphi} := \mathbb{E}_{X\sim\mathbf{p}}[\varphi(X)] = \mathbb{E}_{X\sim\mathbf{g}}\left[\varphi(X)\frac{\mathbf{p}(X)}{\mathbf{g}(X)}\right] \approx \frac{1}{n}\sum_{k=1}^{n}\varphi(X_k)\frac{\mathbf{p}(X_k)}{\mathbf{g}(X_k)} \quad \text{with} \quad X_1,\dots,X_n \sim \mathbf{g}. \tag{1}$$

Generating a sample $X_1,\dots,X_n$ under an importance distribution $\mathbf{g}$ as close as possible to $\mathbf{g}_{\mathsf{opt}} \propto |\varphi| \times \mathbf{p}$ results in a low-variance estimator of $\overline{\varphi}$. A central use case is the rare event setting $\varphi(X) = \mathbb{1}_{X\in\mathbf{F}}$ (with $\overline{\varphi}$ very small) for which Monte Carlo methods are too expansive.

Simulating X according to an efficient importance distribution is notoriously challenging in high dimension. Various approaches have been proposed to address this problem (notably within the MASCOT-NUM community) such as projection into well-chosen subspaces [1, 2] or the use of generative models [3]. The dimension of the space is also an issue when X is not a vector but the trajectory of a stochastic process. This is a common case in reliability assessment where the operation of an industrial system is modeled by a stochastic process (see our previous work [4]).

We propose a new family of importance distributions tailored for dynamic rare event simulation with any non-diffusive Markov process, i.e any piecewise deterministic Markov process (PDMP). These processes evolve according to deterministic differential equations whose parameters are subject to random jumps. Their simulation cost prohibits the learning of a good importance distribution in a non-parametric framework, and the hybrid geometry of their state space prevents the application of most known dimension reduction methods.

The optimal importance distribution for PDMPs was characterized in [5] from the so-called "committor function" of the process. We present a methodology that combines the approximation of this committor function and a cross-entropy procedure to sequentially estimate $\overline{\varphi}$:

1. The PDMP is approximated by a time-homogeneous Markov random walk on a graph.
2. Mean hitting times of a target region of the graph are explicit for the simplified process, and form the basis of a family of approximations of the committor function.

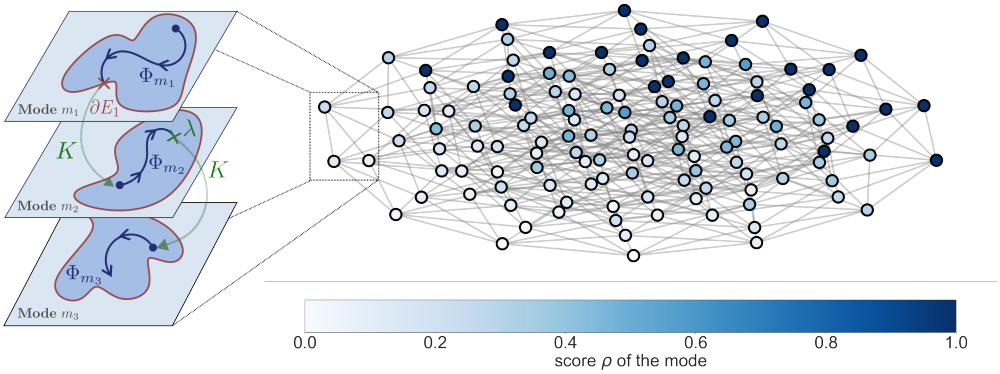


Figure 1: PDMP as a random walk on a graph.

3. This leads to a parametric family of importance distributions, for which we seek a good representative using a cross-entropy method with a new multilevel mechanism.

Our method achieves a variance reduction factor of approximately 10,000 compared to standard Monte Carlo estimation on a large-scale industrial test case.

Short biography (PhD student)

After a master's degree in statistics at Sorbonne University, Guillaume Chennetier did his internship then PhD thesis between École polytechnique and EDF R&D Saclay in rare event simulation for piecewise deterministic Markov processes in order to access the reliability of industrial systems involved in the operation of nuclear and hydraulic power plants.

References

[1] Maxime El Masri. *High dimensional importance sampling through projections on a low dimensional subspace*. PhD thesis, ISAE-Institut Supérieur de l'Aéronautique et de l'Espace, 2022.

[2] Felipe Uribe, Iason Papaioannou, Youssef M Marzouk, and Daniel Straub. Cross-entropy-based importance sampling with failure-informed dimension reduction for rare event simulation. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):818–847, 2021.

[3] Julien Demange-Chryst, François Bachoc, Jérôme Morio, and Timothé Krauth. Variational autoencoder with weighted samples for high-dimensional non-parametric adaptive importance sampling. *arXiv preprint arXiv:2310.09194*, 2023.

[4] Guillaume Chennetier, Hassane Chraibi, Anne Dutfoy, and Josselin Garnier. Adaptive importance sampling based on fault tree analysis for piecewise deterministic markov process. *arXiv preprint arXiv:2210.16185*, 2022.

[5] Hassane Chraibi, Anne Dutfoy, Thomas Galtier, and Josselin Garnier. On the optimal importance process for piecewise deterministic markov process. *ESAIM: Probability and Statistics*, 23:893–921, 2019.

Sampling criteria for excursion set estimation on multi-output models

C. Duhamel [†,1], C. Prieur [§,1], C. Helbert [§,2], M. Munoz Zuniga [§,3], D. Sinoquet [§,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2020 – Jun. 2024*

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK
clement.duhamel@inria.fr ; clementine.prieur@univ-grenoble-alpes.fr

[2] Univ. Lyon, Ecole Centrale de Lyon, CNRS UMR 5208, Institut Camille Jordan,
celine.helbert@ec-lyon.fr

[3] IFP Energies Nouvelles, Rueil-Malmaison
delphine.sinoquet@ifpen.fr ; miguel.munoz-zuniga@ifpen.fr

Abstract

Many industrial issues are related to excursion set estimation problems, formulated as the estimation of a set of feasible black box model input values, that is the set of model input values satisfying a constraint on a model output, for example to remain below a fixed threshold (see for example [2]). An effective way to solve this problem is to model the costly black-box function of interest as a realization of a Gaussian process (GP). This surrogate model is learned thanks to a sequential Design of Experiments, whose points are chosen in the design space $\mathbb{X} \subset \mathbb{R}^d$, accordingly to the optimization of an acquisition criterion (see [5] for more details). The Bichon criterion [1] is a classic criteria for excursion set estimation which provides a good compromise between exploring the design space and exploiting the knowledge around the excursion set boundary.

In this work, we focus on black-box costly models with vectorial output defined by $\boldsymbol{G} := (G_1, \ldots, G_p)$. Partial excursion sets are then defined by :

$$\forall\, i \in \{1, \ldots, p\},\ \Gamma_i^\star := \{\mathbf{x} \in \mathbb{X}, G_i(\mathbf{x}) \leq T_i\}. \tag{1}$$

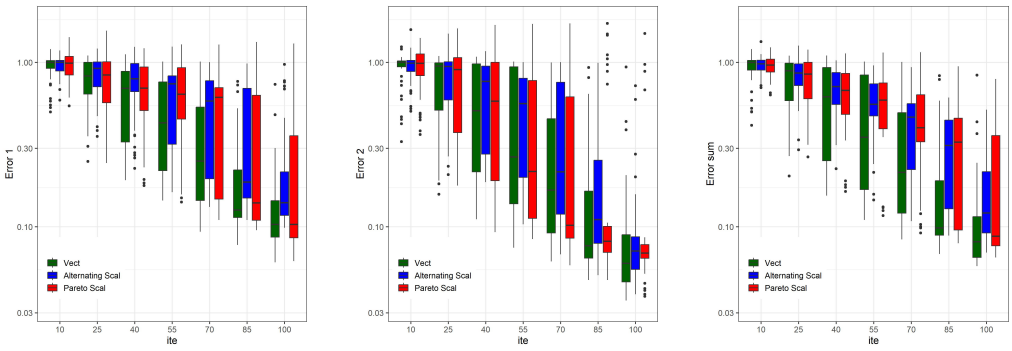
In [3], the authors proposed a criterion suitable for estimating the intersection set of partial excursion sets. In the application which motivates our study, namely a pre-calibration stage of a simulator for the design of floating wind turbines, knowledge of input values feasible for all output components is not sufficient. This is why this work aims to estimate simultaneously each partial excursion sets for each output. It enables to know for a given design space point, which output component exceeds its respective threshold.

In the first part, we introduce two natural multi-dimensional ($p = 2$) approaches based on the one-dimensional Bichon criterion. The first approach (called *Alternating Scal*) optimizes each Bichon criterion alternatively between the two components. The second approach (called *Pareto Scal*) is associated with some Pareto solutions of the bi-objective optimization problem of Bichon criteria. These two procedures use as many classic GP models as outputs (one for each component). Then, we propose a vectorial extension (called *Vect*) of the Bichon criterion based

*I benefit from an adaptation of the thesis (part-time 80% + extension of the contract) for health reasons.

on the minimum of the distances between each component of the GP and its corresponding threshold. This extension is naturally based on a multi-output GP model which takes into account the correlation between the outputs (see [4] for details) and requires calculation of orthants probabilities of multivariate normal distributions.

The different methodologies proposed above are compared on several analytical examples (with 2 then 4 input components and 2 output components), including the example below of the augmented Hartmann 4d function with 2 components and their associated thresholds $T = (-1, -1.6)$. Work on implementing these methods on the target application (application to the pre-calibration of a simulator for the design of floating wind turbines) is in progress.



(a) $\frac{\text{Vol}(\hat{\Gamma}_1 \Delta \Gamma_1^\star)}{\text{Vol}(\Gamma_1^\star)}$ (b) $\frac{\text{Vol}(\hat{\Gamma}_2 \Delta \Gamma_2^\star)}{\text{Vol}(\Gamma_2^\star)}$ (c) $\sum_i \frac{\text{Vol}(\hat{\Gamma}_i \Delta \Gamma_i^\star)}{\text{Vol}(\Gamma_i^\star)}$

Figure 1: Boxplots of partial relative errors (and sum), for the different criteria, in the case of 40 init. DoE of size 20 with 100 iterations, for the augmented Hartmann 4d function with $T = (-1.6, -1)$.

Short biography (PhD student)

I'm a fourth year PhD student in AIRSEA team of Laboratoire Jean Kuntzmann at Univ. Grenoble Alpes. My thesis project initiated in October 2020 is about sampling criteria for solving scalar or vectorial excursion set estimation problems using Gaussian processes, with application to wind turbines simulator pre-calibration. This project is funded by INRIA, and is part of a collaboration with IFPEN.

References

[1] Barron J Bichon, Michael S Eldred, Laura Painton Swiler, Sandaran Mahadevan, and John M McFarland. Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA journal*, 46(10):2459–2468, 2008.

[2] Mohamed Reda El Amri, Céline Helbert, Olivier Lepreux, Miguel Munoz Zuniga, Clémentine Prieur, and Delphine Sinoquet. Data-driven stochastic inversion via functional quantization. *Statistics and Computing*, 30:525–541, 2020.

[3] Trygve Olav Fossum, Cédric Travelletti, Jo Eidsvik, David Ginsbourger, and Kanna Rajan. Learning excursion sets of vector-valued gaussian random fields for autonomous ocean sampling. *The annals of applied statistics*, 15(2):597–618, 2021.

[4] Haitao Liu, Jianfei Cai, and Yew-Soon Ong. Remarks on multi-output gaussian process regression. *Knowledge-Based Systems*, 144:102–121, 2018.

[5] Victor Picheny, David Ginsbourger, Olivier Roustant, Raphael T Haftka, and Nam-Ho Kim. Adaptive designs of experiments for accurate approximation of a target region. 2010.

A stepwise uncertainty reduction strategy for the estimation of small quantile sets

Romain Ait Abdelmalek-Lomenech[†,1], Julien Bect[‡,1], Emmanuel Vazquez[‡,1]

[†] PhD student (presenting author). [‡] PhD supervisor

PhD expected duration: Oct. 2021 – Sep. 2024

[1] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, 91190 Gif-sur-Yvette, France
{romain.ait, julien.bect, emmanuel.vazquez}@centralesupelec.fr

Abstract

Given a numerical simulator of a physical phenomenon or system, one often seeks to determine the set of inputs that lead to values with specified properties. Such problems are broadly known as *set inversion* problems.

This communication focuses on a particular robust formulation of such problems that we call *Quantile Set Inversion* (QSI) [1], in which the function of interest has both deterministic and uncertain inputs. In this formulation, given a vector-valued function $f : \mathbb{X} \times \mathbb{S} \mapsto \mathbb{R}^q$, a critical region $C \subset \mathbb{R}^q$, and a threshold $\alpha \in (0, 1)$, the object of interest is the set

$$\Gamma(f) = \{x \in \mathbb{X} \ : \ \mathbb{P}(f(x, S) \in C) \leq \alpha\},$$

where $\mathbb{X}$ and $\mathbb{S}$ are respectively the spaces of design and uncertain input variables, and S is a random vector with known probability distribution $\mathbb{P}_S$.

This object, which we call a *quantile set*, corresponds to the set of deterministic inputs such that the probability (w.r.t. the distribution of the uncertain inputs) that the output variables belong to a critical region is below a threshold. A simple scalar-valued example of such a problem is shown in Figure 1.

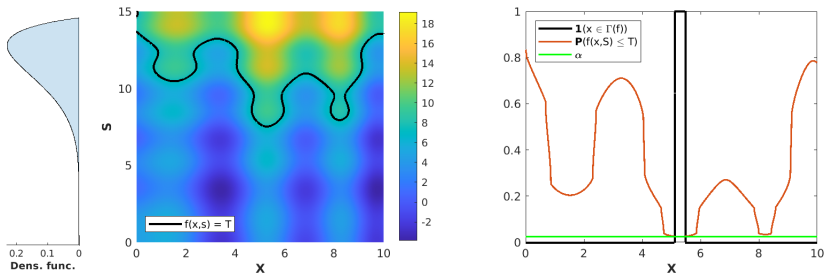


Figure 1: Scalar-valued example function (middle), probability density function associated to the distribution $\mathbb{P}_S$ of the uncertain input variable (left), and set of interest obtained with critical region $C = (-\infty, T]$ and threshold $\alpha = 0.025$ (right).

In the case of an *expensive-to-evaluate* function, the objective is to reach a good approximation of the set of interest using only a small, predetermined number of evaluations. The choice of the evaluation points must then be conducted carefully.

To address this set inversion problem, while taking into consideration the expensive-to-evaluate nature of the underlying function, we have proposed a sequential Bayesian strategy [1] based on the *Stepwise Uncertainty Reduction* (SUR) principle (see, e.g., [5] and references therein). Starting from a Gaussian process prior on the unknown function f, the strategy selects the next evaluation points by minimizing the expected future uncertainty about the *quantile set*, measured by a particular metric inspired by [3].

We propose, in this communication, an improvement of the method discussed in [1] to tackle more difficult problems where the quantile set is small with respect to the full input domain $\mathbb{X}$. In particular, we focus on the resolution of two defects arising in such context:

- An adaptive sequential Monte Carlo algorithm is introduced, in the spirit of [4], to tackle the difficulty of the sampling criterion optimization in the case of small quantile sets.
- The criterion is estimated using approximated Gaussian process sample paths [2], allowing an improvement in term of computation cost.

The performance of this strategy is illustrated by applying it to several test functions.

Short biography (PhD student)

With a backgroung in probability theory, statistics, machine learning and quantitative finance, Romain Ait Abdelmalek-Lomenech began his PhD in October 2021 with CentraleSupélec, under the supervision of Emmanuel Vazquez and Julien Bect. His work aims at developing new Bayesian methods for the optimization/inversion of expensive-to-evaluate functions in presence of uncertain input variables.

The present work is part of a PhD thesis funded by the French National Research Agency (ANR) in the context of the SAMOURAI project (ANR-20-CE46-0013).

References

[1] R. Ait Abdelmalek-Lomenech, J. Bect, V. Chabridon, and E. Vazquez. Bayesian sequential design of computer experiments for quantile set inversion. *arXiv preprint arXiv:2211.01008v3*, 2023.

[2] D. Azzimonti, J. Bect, C. Chevalier, and D. Ginsbourger. Quantifying uncertainties on excursion sets under a Gaussian random field prior. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):850–874, 2016.

[3] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3): 773–793, 2012.

[4] J. Bect, L. Li, and E. Vazquez. Bayesian Subset Simulations. *SIAM/ASA Journal on Uncertainty Quantification*, 5:762–786, 2017.

[5] J. Bect, F. Bachoc, and D. Ginsbourger. A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919, 2019.

High-Dimensional Bayesian Optimization with a Combination of Kriging models

Tanguy Appriou[†,1,2], Didier Rullière[§,2], David Gaudrie[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2021 – Nov. 2024

[1] Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France
tanguy.appriou@emse.fr
drulliere@emse.fr

[2] Stellantis, Centre d'Expertise R&D Carrières-sous-Poissy, F-78955 Carrières-sous-Poissy, France
david.gaudrie@stellantis.com

Abstract

Kriging models, also called Gaussian Process regression, [7] are commonly used as surrogate models of expensive computer codes in various applications. In engineering design optimization, Bayesian approaches based on Kriging models such as efficient global optimization (EGO) [6] are employed to speed-up the optimization process by reducing the number of function evaluations. These methods have been successfully applied to many real-world applications in low dimensions (less than 30 design parameters) [4].

However, engineering designs can often be parametrized by more than 50 parameters in practice. In higher dimensions, Kriging suffers from the curse of dimensionality and building an accurate surrogate model is met with various setbacks. One of the main challenges is related to optimizing the covariance length-scale hyperparameters of the model. These hyperparameters regulate the decay of the correlation between observations when their distance increases, and anisotropic Kriging models consider one length-scale per dimension. Estimating the length-scales correctly is essential to obtain a model with a good accuracy. They are typically obtained by maximizing the log-likelihood of the model. In high-dimension, this inner optimization is problematic due to the exponential growth of the search space with the dimension and to over-fitting issues when there are too few observations. For these reasons, maximum likelihood estimation of the hyperparameters often fails to provide correct values, especially when not enough observations are available [5, 1]. This is problematic for Bayesian optimization where the number of samples at the start of the optimization is low, and where inaccurate surrogates negatively impact the convergence speed. Several papers address these difficulties by either reducing the dimension of the problem [2], or by considering simplifying hypothesis such as additive models [3]. However, both these methods make additional assumption on the underlying function to approximate (low dimension representation or additive structure) which are not necessarily satisfied in practice, thus they are not easily generalizable to any design engineering problems.

In this presentation, we introduce a new method for high-dimensional Bayesian optimization using a convex combination of Kriging models. This method bypasses the length-scales optimization by combining sub-models with random length-scales and its expression is obtained in closed-form expression avoiding any inner optimization. We also describe how to sample suitable length-scales for the sub-models using a criterion based on the entropy of the correlations,

in order to avoid degenerated sub-models with either too large or too small length-scales. Finally, the variance of the combination being not directly available as the correlations between sub-models are unknown, we present a method to compute the prediction variance for any weighting methods. This is done by introducing a global covariance structure also based on a linear combination of covariances. We apply our combined Kriging model to high-dimensional EGO for analytical test functions and for the design of an electric machine. We show that the classical Kriging approach using maximum likelihood estimation fails to properly optimize the length-scale hyperparameters and that our method successfully builds more accurate surrogate models at the beginning of the optimization loop. This results in faster convergence speed for EGO using the combination, thus reducing the number of computer code evaluations to reach the optimum.

Short biography (PhD student)

Tanguy Appriou received an engineering degree from Ecole Centrale de Lyon and a master in aerospace engineering from Tohoku University in 2021. He is currently pursuing a Ph.D. degree in Stellantis and Mines Saint-Etienne. This thesis is partly funded by a CIFRE grant (convention #2021/1284) established between the ANRT and Stellantis.

This thesis is conducted with the support of the consortium in Applied Mathematics CIROQUO (`https://doi.org/10.5281/zenodo.6581217`), gathering partners in technological and academia in the development of advanced methods for Computer Experiments.

References

[1] Tanguy Appriou, Didier Rullière, and David Gaudrie. Combination of optimization-free kriging models for high-dimensional problems. *Computational Statistics*, pages 1–23, 2023.

[2] Paul G Constantine. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM, 2015.

[3] Nicolas Durrande, David Ginsbourger, and Olivier Roustant. Additive Covariance kernels for high-dimensional Gaussian Process modeling. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 21(3):481–499, 2012.

[4] Alexander Forrester, Andras Sobester, and Andy Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.

[5] David Ginsbourger, Delphine Dupuy, Anca Badea, Laurent Carraro, and Olivier Roustant. A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments. *Applied Stochastic Models in Business and Industry*, 25(2):115–131, 2009.

[6] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[7] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

Quantifying uncertainties in seismic waves propagation with a Fourier Neural Operator surrogate model

F. Lehmann[†,1,2], F. Gatti[§,2], M. Bertin[§,1], D. Clouteau[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2021 – Sep. 2024

[1] CEA/DAM/DIF, Arpajon, France
michael.bertin@cea.fr

[2] LMPS - Laboratoire de Mécanique Paris-Saclay, Université Paris-Saclay, CentraleSupélec, ENS Paris-Saclay, CNRS, Gif-sur-Yvette, France
{fanny.lehmann,filippo.gatti,didier.clouteau}@centralesupelec.fr

Abstract

Physics-based numerical simulations are key tools in earthquake engineering. They complement the existing datasets of recorded earthquakes by computing on-demand the ground motion generated by any realistic earthquake scenario. However, high-fidelity earthquake simulations are computationally demanding since they require solving the hyperbolic elastic wave equation in large three-dimensional (3D) domains and up to high frequencies.

In addition, simulation parameters are highly uncertain due to the difficulty of conducting geophysical experiments. Parameters of particular interest are i) the ground properties that control the waves velocity, ii) the position of the earthquake source, iii) the properties of the earthquake source (i.e. orientation and magnitude). Due to the cost of numerical simulations, repeated calls to the numerical solvers are unaffordable and efficient surrogate models are required to quantify uncertainties. Existing surrogate models based on e.g. Gaussian processes [1] or Polynomical Chaos Expansion [5] do not allow 3D applications with highly heterogeneous domains.

In this work, we propose a surrogate model of seismic waves propagation using a Factorized Fourier Neural Operator (F-FNO [6]), a deep learning method tailored to Partial Differential Equations (PDEs). The F-FNO views integral operators as convolutional kernels of learnable weights and writes the convolution as a product of Fourier coefficients. This leads to efficient neural operators that have strong relationships with physical equations.

Our training database is built from our HEMEW-3D database of 30,000 High-Performance Computing (HPC) simulations [2]. For each simulation, a heterogeneous propagation domain is designed with 3D random fields that represent variations of the rock properties inside the ground (Fig. 1, left). The source position and orientation are also randomly chosen for each simulation. Then, seismic waves are propagated from the source up to the surface where they are recorded by a grid of virtual sensors. Therefore, for each set of input parameters (ground properties, source position, source orientation), the F-FNO learns to predict the time-dependent surface wavefields (Fig. 1).

The F-FNO is compared to baseline models and we show that it is an efficient surrogate model whose accuracy improves when the network complexity increases [3]. Prediction errors are also quantified in the frequency domain, which indicates that the well-known spectral bias hinders

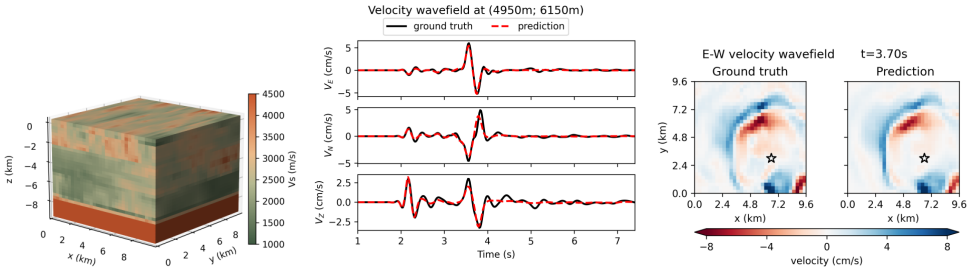


Figure 1: For a heterogeneous geology (left), timeseries of surface ground motion (middle) and the surface wavefields at $t = 3.7$s (right). F-FNO predictions (dashed red lines, middle) are compared with the outputs of the numerical simulations (black lines, middle).

high-frequency accuracy. In addition, prediction errors are well explained by properties of the inputs, thereby giving insights on the expected accuracy before making the prediction.

To quantify the influence of geological uncertainties on surface wavefields, transfer learning was applied to a small dataset of real geologies in Southeastern France [4]. Thanks to the (almost) negligible cost of the F-FNO evaluation, once duly trained, we obtained meaningful confidence intervals that are of great significance for the earthquake engineering community.

Short biography (PhD student)

I graduated from the mathematics department of the ENS Paris-Saclay in 2021. I am doing my PhD at CEA DAM, which partners with academic institutions like Centrale Supélec to perform seismic hazard analyses. This PhD builds on several years of development of the HPC code SEM3D to design a meta-model enabling fast and accurate predictions of seismic ground motion. The meta-model will then be used to investigate site effects for seismic hazard analyses.

References

[1] A. Alimoradi and J. L. Beck. “Machine-Learning Methods for Earthquake Ground Motion Analysis and Simulation”. In: *Journal of Engineering Mechanics* 141.4 (Apr. 2015), p. 04014147. DOI: 10.1061/(ASCE)EM.1943-7889.0000869.

[2] F. Lehmann et al. “Synthetic Ground Motions in Heterogeneous Geologies: The HEMEW-3D Dataset for Scientific Machine Learning”. In: *Earth System Science Data Discussions* 2024 (2024), pp. 1–26. DOI: 10.5194/essd-2023-470.

[3] F. Lehmann et al. “3D Elastic Wave Propagation with a Factorized Fourier Neural Operator (F-FNO)”. In: *Computer Methods in Applied Mechanics and Engineering* 420 (2024), p. 116718. DOI: 10.1016/j.cma.2023.116718.

[4] F. Lehmann et al. “Seismic Hazard Analysis with a Factorized Fourier Neural Operator (F-FNO) Surrogate Model Enhanced by Transfer Learning”. In: *NeurIPS 2023 AI for Science Workshop*. 2023.

[5] P. Sochala, F. De Martin, and O. Le Maître. “Model Reduction for Large-Scale Earthquake Simulation in an Uncertain 3D Medium”. In: *International Journal for Uncertainty Quantification* 10.2 (2020), pp. 101–127. DOI: 10.1615/Int.J.UncertaintyQuantification.2020031165.

[6] A. Tran et al. “Factorized Fourier Neural Operators”. In: *The Eleventh International Conference on Learning Representations*. 2023.

Variational autoencoder with weighted samples for high-dimensional non-parametric adaptive importance sampling

Julien Demange-Chryst[†,1,2], François Bachoc[§,2], Jérôme Morio[§,1], Timothé Krauth[1,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2021 – Sep. 2024

[1] ONERA/DTIS, Université de Toulouse, F-31055 Toulouse, France
{julien.demange-chryst,jerome.morio}@onera.fr

[2] Institut de Mathématiques de Toulouse, UMR5219 CNRS, 31062 Toulouse, France
francois.bachoc@math.univ-toulouse.fr

[3] Zurich University of Applied Sciences, Centre for Aviation, Winterthur, Switzerland
timothe.krauth@zhaw.ch

Abstract

Importance sampling (IS) is a well-known uncertainty quantification method, classically used as a variance-reduction technique for Monte Carlo integration including rare event estimation [6], or for generating points from a target probability distribution known up to a constant [5]. The common denominator of every importance sampling procedure is that they all require to estimate a target probability distribution with weighted samples, and obviously, the accuracy of the algorithm depends on the quality of the estimation of the distribution. Moreover, we also need to be able to not only sample from the built auxiliary distribution, but also to have access to its PDF values.

A first way to perform this density estimation is to use non-parametric models, such as kernel smoothing. These models are flexible, but despite some improvements they strongly suffer from the curse of dimensionality since the size of the required sample to have a good approximation of the target distribution exponentially grows with the dimension. Another solution is to use parametric families of distributions, such as the Gaussian or Gaussian mixture ones, which are more robust in medium-high dimension. However, they sometimes require some prior knowledge on the target distribution, and their lack of flexibility and the huge number of parameters to estimate can negatively impact the quality of the estimation when the dimension is high.

In order to combine both flexibility and robustness faced to the dimension, we propose to use as the auxiliary sampling distribution a distribution parameterised by a variational autoencoder [4], whose main principle has been introduced in the last decade. Variational autoencoders are deep generative models for approximating high-dimensional complex distributions of observed data and generating new samples. The specific feature of a variational autoencoder compared to other density estimation methods is that it performs a dimensionality reduction into a lower dimensional latent space in order to facilitate the estimation. Moreover, in opposition to other dimensionality reduction techniques such as principal component analysis or autoencoders, variational autoencoders have good generation properties and give explicitly the approximating distribution, allowing to perform Monte Carlo simulations. This tool is now popular in the machine learning community but not so much in uncertainty quantification.

In the present communication [3], we extend the existing framework of variational autoencoders to the case of weighted samples by introducing a new objective function. The resulting IS auxiliary

distribution is close to an infinite mixture of Gaussian distributions. Then, its flexibility makes it as expressive as a non-parametric model, and despite the very high number of parameters to estimate, it is much more efficient in high dimension than the classical Gaussian or Gaussian mixture families. Moreover, in order to add even more flexibility to the model and to be able to learn multimodal distributions, we consider a learnable prior distribution for the variational autoencoder latent variables. We also introduce a new pre-training procedure for the variational autoencoder to find good starting weights of the neural networks to prevent as much as possible the posterior collapse phenomenon to happen.

At last, we explicit how the resulting distribution can be combined with importance sampling. Indeed, the existing procedure [7] to compute the PDF values of the resulting distribution of a variational autoencoder leads to a biased and non-convergent importance sampling estimator. In order to keep an unbiased and consistent estimator, we introduce a new way to compute the PDF values. Then, we show how to integrate the whole suggested procedure into existing reliability algorithms, such as the cross-entropy algorithm, for rare event estimation. Finally, we illustrate the practical interests of the previous efforts on two multimodal rare-event-estimation problems. The code to reproduce the numerical experiments is publicly available at: `https://github.com/Julien6431/Importance-Sampling-VAE.git`.

Short biography

I graduated in 2021 from the engineering school ISAE-SUPAERO in Toulouse and I also obtained a MSc degree in applied mathematics from Toulouse III - Paul Sabatier University. Then, I performed my final-year internship at *ONERA* in Toulouse which led to my current PhD thesis co-funded by *ONERA* and *Toulouse III - Paul Sabatier University*. My first two topics of research of my PhD on reliability-oriented sensitivity analysis with dependant inputs and the common estimation of multiple expectations lead to two journal publications [2, 1].

References

[1] Julien Demange-Chryst, François Bachoc, and Jérôme Morio. Efficient estimation of multiple expectations with the same sample by adaptive importance sampling and control variates. *Statistics and Computing*, 33(5):103, 2023.

[2] Julien Demange-Chryst, François Bachoc, and Jérôme Morio. Shapley effect estimation in reliability-oriented sensitivity analysis with correlated inputs by importance sampling. *International Journal for Uncertainty Quantification*, 13(3), 2023.

[3] Julien Demange-Chryst, François Bachoc, Jérôme Morio, and Timothé Krauth. Variational autoencoder with weighted samples for high-dimensional non-parametric adaptive importance sampling. *arXiv preprint arXiv:2310.09194*, 2023.

[4] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *Article accepted in the 2nd International Conference on Learning Representations 2014*, 2014.

[5] Jean-Michel Marin, Pierre Pudlo, and Mohammed Sedki. Consistency of adaptive importance sampling and recycling schemes. *Bernoulli*, 25(3):1977 – 1998, 2019.

[6] Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte Carlo simulation, and machine learning*, volume 133. Springer, 2004.

[7] Hechuan Wang, Mónica F Bugallo, and Petar M Djurić. Adaptive importance sampling supported by a variational auto-encoder. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 619–623. IEEE, 2019.

Sensitivity analysis of set-valued models

N. Fellmann [†,1,2], C. Helbert[§,1], C. Blanchet[§,1], A. Spagnol[§,2], D. Sinoquet[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Dec. 2021 – Nov. 2024

[1] École Centrale de Lyon, CNRS UMR 5208, Institut Camille Jordan
{noe.fellmann,christophette.blanchet,celine.helbert}@ec-lyon.fr

[2] IFP Energies Nouvelles
{noe.fellmann,delphine.sinoquet,adrien.spagnol}@ifp.fr

Abstract

Sensitivity analysis (SA) plays a central role in mathematical modeling, providing valuable insight into how input parameters variations affect a model output. This tool allows practitioners to evaluate the robustness and reliability of their models by quantifying the response to changes in key variables. Historically, SA methods have been tailored for scalar-valued models. However, some physical problems involve more complex outputs, such as set-valued models, where each model evaluation represents a subspace within a larger space $\mathcal{X}$. Such models occur in several domains, for example, when the output of interest is the geographic area where a measured pollutant concentration exceeds a specified threshold (Figure 1). It becomes critical to identify which input parameters (e.g., traffic or environmental factors) significantly affect the polluted area. Traditional SA methods are not directly applicable or adaptable to set-valued models, posing a challenge in quantifying uncertainties in the set-valued response.

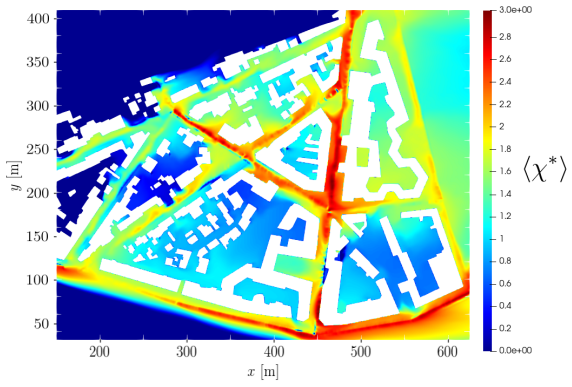


Figure 1: Concentration maps of pollutant dispersion at the urban scale

To fill this gap, we propose three different approaches for defining sensitivity indices when dealing with set-valued models. Our first two are Sobol-like indices S_i^V based on random set tools [3] and universal indices S_i^{Univ} [2] adapted to set outputs. The third approach uses kernel-based sensitivity indices, specifically HSIC and HSIC-ANOVA [1], with sets. This requires the introduction of a kernel, denoted k_{set}, defined on sets, which we thoroughly investigate. In

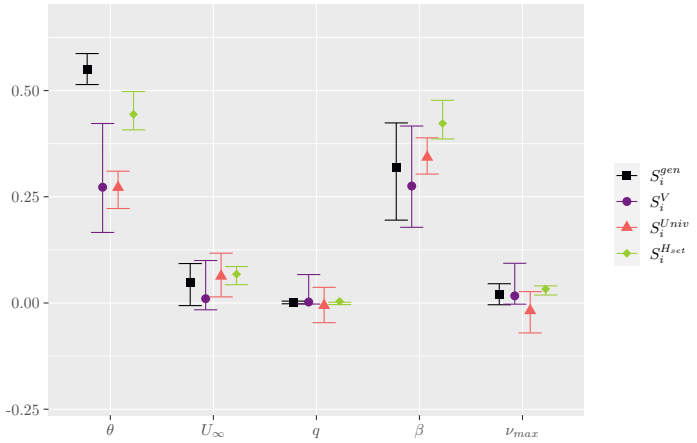


Figure 2: Comparison of the four indices on the set-valued model of pollutant concentration maps. 1000 model evaluations are used to estimate indices and confidence intervals

particular, we demonstrate its *characteristic* property, based on a result of [4], an essential property within the HSIC framework. k_{set} is then used to estimate HSIC with set-valued outputs, but this estimation is complicated by the presence of sets, which forces k_{set} to be estimated as well. Therefore, we propose and study a nested Monte Carlo estimator of the indices.

In Figure 2, our three new indices S_i^V, S_i^{Univ}, and $S_i^{H_{set}}$ are compared to the generalized point by point Sobol index S_i^{gen} on the pollutant concentration maps. The SA results are similar, but $S_i^{H_{set}}$ stands out as it has less variability than the other three and has the strength to be used for both screening and ranking.

Another ongoing application is the use of these indices in chance-constrained problems and robust optimization by quantifying the impact of uncertain inputs on excursion sets. Knowing which uncertain inputs affect the optimization could allow simplifying steps within Bayesian optimization.

Short biography

I am in the third year of my PhD, which is taking place at the Ecole Centrale de Lyon and IFP Energies Nouvelles. My work is about using sensitivity analysis methods to simplify robust optimization problems. I'm supervised by Céline Helbert, Christophette Blanchet from ECL and Adrien Spagnol, Delphine Sinoquet from IFPEN. My thesis is part of the CIROQUO consortium.

References

[1] Sébastien Da Veiga. Kernel-based ANOVA decomposition and Shapley effects - Application to global sensitivity analysis. working paper or preprint, January 2021.

[2] Jean-Claude Fort, Thierry Klein, and Agnès Lagnoux. Global sensitivity analysis and wasserstein spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):880–921, 2021.

[3] Ilya Molchanov. *Theory of random sets*, volume 19. Springer, 2005.

[4] Johanna Ziegel, David Ginsbourger, and Lutz Dümbgen. Characteristic kernels on hilbert spaces, banach spaces, and on sets of measures, 2022.

I-optimal sequential design for Bayesian inverse problems with Gaussian process surrogate models

P. Lartaud[†,1,2], P. Humbert[2], J. Garnier[1],

PhD expected duration: Oct. 2021 – Sep. 2024

[1] CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
{paul.lartaud,josselin.garnier}@polytechnique.edu

[2] CEA DAM-DIF, Bruyères-le-Châtel 91297 Arpajon, France
philippe.humbert@cea.fr

Abstract

In the context of nuclear safeguards and security, passive neutron correlation measurements can be used to identify an unknown fissile matter [2]. This approach has been widely used and documented but the focus on the uncertainty quantification is more recent. The common approach is to solve an inverse problem with a Bayesian approach, while building a surrogate model to overcome the computational burden of the true direct model [3]. The training data for the surrogate can be obtained using state-of-the-art neutronic Monte-Carlo codes, but the limited computational budget restricts the number of data points that can be obtained. As a consequence, a large portion of the uncertainties in the inverse problem may reside in the lack of knowledge of the surrogate itself. Hence the problem at stake: given some observations of an inverse problem, how can we refine the surrogate model in the region of interest by feeding it with well-chosen design points ? To answer this question, we developed an I-optimal stepwise uncertainty reduction (ISUR) strategy for the sequential design of computer experiments in the context of Bayesian inverse problems with Gaussian process surrogate models. This method is based on the Stepwise Uncertainty Reduction (SUR) paradigm introduced in [1] and [4]. Let $\mathbb{M}$ be the space of Gaussian measures on an appropriate functional space. We consider an inverse problem with noisy observations $\mathbf{y} \in \mathbb{R}^N$ and we are trying to identify the posterior distribution $p(x|\mathbf{y})$ for some parameters $x \in \mathcal{X} \subset \mathbb{R}^p$. The newly proposed method is derived from I-optimal designs with a metric of uncertainty given by the function $\mathcal{H}\colon \mathbb{M} \to [0, +\infty)$ where the functional is defined for $\nu \in \mathbb{M}$ by:

$$\mathcal{H}(\nu) = \mathbb{E}_{p_\nu}\left[k_\nu(\tilde{x}, \tilde{x})\right] = \int_{\mathcal{X}} k_\nu(\tilde{x}, \tilde{x}) p_\nu(\tilde{x}|\mathbf{y}) d\tilde{x} \tag{1}$$

where $k_\nu(\tilde{x}, \tilde{x})$ is the predictive variance on $\tilde{x}$ of the GP associated with the measure ν and $p_\nu(\tilde{x}|\mathbf{y})$ is the posterior distribution obtained from noisy observations $\mathbf{y}$ and using the GP surrogate model as a direct model.
This metric can be understood as the integrated predictive variance of the GP surrogate over the posterior distribution of the inverse problem. We show that the ISUR criteria for this particular metric can be derived analytically and we verify the supermartingale property and the almost-sure convergence of the metric towards 0. The results can also be extended to a multi-dimensional framework in which $\mathbf{y} \in \mathbb{R}^{N \times d}$. Finally we also introduced a simpler D-optimal design strategy and compare the two methods with a naive design on a number of test cases.

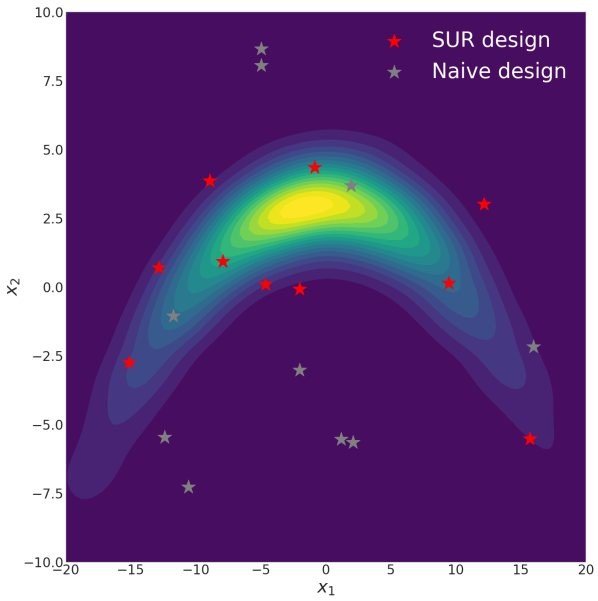


Figure 1: New design points with naive (grey) and ISUR (red) strategies on an analytical inverse problem test case with a banana-shaped posterior distribution

Short biography (PhD student)

I am a third-year PhD student working at CEA DAM Île-de-France and École polytechnique under the supervision of Josselin Garnier. This work is funded by the CEA and aims at providing robust uncertainty quantification methods for the identification of nuclear materials using passive neutron correlation measurements.

References

[1] Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22:773–793, 2012.

[2] Richard Phillips Feynman, Frederic De Hoffmann, and Robert Serber. Dispersion of the neutron emission in U-235 fission. *Journal of Nuclear Energy (1954)*, 3(1-2):64–IN10, 1956.

[3] Paul Lartaud, Philippe Humbert, and Josselin Garnier. Multi-output gaussian processes for inverse uncertainty quantification in neutron noise analysis. *Nuclear Science and Engineering*, 197(8):1928–1951, 2023.

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

Quantifying the uncertainties of a multidimensional and multifidelity nonlinear simulation, application to the calculation of the deformation of fuel assemblies in a pressurized water reactor

A. Abboud[†,1,2], J. Garnier[§,1], B. Leturcq[§,2], S. de Lambert [§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2022 – Oct. 2025

[1] CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
{ali.abboud, josselin.garnier}@polytechnique.edu

[2] Université Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, 91191, Gif-sur-Yvette, France
{ali.abboud,bertrand.leturcq, Stanislas.DE-LAMBERT}@cea.fr

Abstract

In the core of nuclear reactors, fluid-structure interaction [1] and intense irradiation lead to progressive deformation of fuel assemblies. When this deformation is significant, it can lead to additional costs and longer fuel unloading and reloading operations.

Therefore, it is preferable to adopt a fuel management that avoids excessive deformation and interactions between fuel assemblies. However, the prediction of deformation and interactions between fuel assemblies is uncertain. Uncertainties affect neutronic, thermal-hydraulic and thermomechanical parameters. Indeed, the initial uncertainties are propagated over several successive power cycles of twelve months each through the coupling of nonlinear, nested and multidimensional thermal-hydraulic and thermomechanical simulations.

In this work, we set out to study the hydraulic contribution and quantify the associated uncertainty. To achieve this objective, we develop a multi-stage approach to carry out an initial sensitivity analysis [4], highlighting the most influential parameters in the hydraulic model. By optimally adjusting these parameters, we aim to obtain a more accurate description of the flow redistribution phenomenon in the reactor core. The aim of the sensitivity analysis presented in this work is to construct an accurate and suitable surrogate model [3] that represents the flow redistribution in the core. This surrogate model will then be coupled with the thermomechanical model to quantify the final uncertainty in the simulation of fuel assembly deformation within a pressurised water reactor. This approach will provide a better understanding of the interactions between hydraulic and thermomechanical phenomena, thereby improving the reliability and accuracy of the simulation results.

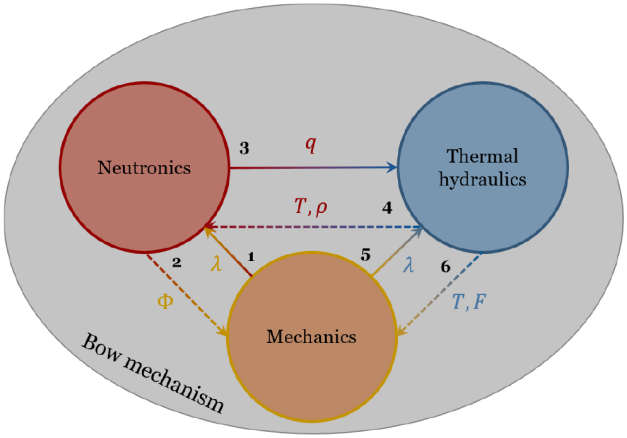


Figure 1: Physical interactions during the deformation of the assembly (From [2])

Short biography (PhD student)

After completing a master's degree in mathematics and applications at the University of Reims Champagne-Ardenne, I did an internship in finite elements applied to neutronics in the Reactor Studies and Applied Mathematics Department (SERMA) at CEA Saclay. I then went on to do a PhD at the École Polytechnique, funded by the French Atomic Energy Commission (CEA), in the field of statistics and probability applied to the deformation of fuel assemblies in a nuclear reactor, in the Mechanical and Thermal Studies Department (SEMT).

References

[1] Wanninger A. *"Mechanical Analysis of the Bow Deformation of Fuel Assemblies in a Pressurized Water Reactor Core"*. Dissertation, Technische Universität München, München, 2018.

[2] S. De Lambert. *"Contribution to the multiphysical analysis of fuel assembly bow"*. PhD thesis, 2021. 2021UPAST035.

[3] C. E. Rasmussen and Christopher K. I. Williams. *"Gaussian Processes for Machine Learning"*, volume 2. MIT Press, Cambridge, MA, 2006.

[4] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. "Sensitivity Analysis in Practice : A Guide to Assessing Scientific Models".

Optimal Design of Physical and Numerical Experiments for Computer Code Calibration

Adama BARRY[†,1,2,3],
François BACHOC[§,1], Clémentine PRIEUR[§,2],
Sarah BOUQUET[3], Miguel Munoz ZUNIGA[3]

[†] PhD student. [§] PhD supervisor

PhD expected duration: Jan. 2022 - Jan. 2025

[1] Institut de Mathématiques de Toulouse

[2] Université de Grenoble, INRIA Grenoble

[3] IFP Energies Nouvelles

Abstract

In industry, the development of computer codes is often implemented to study and analyze complex physical phenomena or systems. Some of these computer codes depend on two types of variables: experimental or control variables and intrinsic parameters. These parameters are often physical constants and/or control parameters, which have no physical interpretation. Precise values must be set for these parameters by the engineer so that the computer code imitates the physical phenomenon or system of interest as closely as possible. The complex nature of these computer codes requires an efficient calibration process, in which unknown parameters are adjusted to improve alignment between computer code outputs and observations of physical phenomena. This calibration process is essential to ensure the accuracy and reliability of the expensive computer code, enabling a more informed understanding of the phenomenon or system of interest. Most of the work on Bayesian calibration focuses on building the computer code emulator by selecting the computer simulations to be carried out without worrying about the quality of the physical measurements. Since these measurements are limited by their cost or the difficulty of acquiring them, it would be wise to select them for more effective calibration.

The selection of optimal designs of physical and numerical experiments in the calibration procedure plays a crucial role in achieving robust calibration using Bayesian methods. Based on the classical Bayesian framework of [6], we propose a hybrid algorithm for selecting the design of physical experiments and the design of numerical experiments to solve the calibration problem. This strategy enables precise construction of the computer code emulator, leading to a better approximation of the posterior density of the calibration parameters, and consequently, to more accurate calibration results.

The first step is to select the physical experiments. We will therefore begin by presenting criteria for measuring the quality of a design of physical experiments. These criteria can be grouped into two categories: those based on the information matrix [7, 4] and those based on the exact a posteriori distribution. The latter are better suited to the calibration problem because they take into account the non-linear nature of the computer code, the uncertainty in the physical phenomenon, and the calibration parameters. However, they are expensive to evaluate due to the use of Monte Carlo procedures. The first challenge is the rapid evaluation of these criteria, and the second is solving the resulting optimization problem. For the first, we will present a fast

calculation method without a Markov chain Monte Carlo procedure and a variant of simulated annealing for the second.
The second step consists of selecting the design of numerical experiments. We will present criteria from the literature [3, 2] and those we propose. Our criteria are inspired by the sequential uncertainty reduction paradigm (SUR) [1]. They are based on uncertainty measurements for the calibration objective. For their optimization, which is costly, we will use a greedy algorithm that exploits the Monte Carlo procedure used in their computation. A comparative study on a toy case will be presented at the end to illustrate the performance of the different algorithms.

Short biography (PhD student)

Graduate of ENSAI in Data Science and INSEA in Statistics and Applied Economics. I am currently in the 3rd year of my PhD and am affiliated with IFP Energies Nouvelles, the Institut de Mathématiques de Toulouse, and INRIA Grenoble. I'm working on the Bayesian calibration of expensive computer code. The thesis project is funded by IFPEN as part of CIROQUO (Consortium Industrie Recherche pour l'Optimisation et la QUantification d'incertitude pour les données Onéreuses).

References

[1] Clément Chevalier, David Ginsbourger, Victor Picheny, Julien Bect, Emmanuel Vazquez, and Yann Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.

[2] Xiaowu Dai and Peter Chien. Another look at statistical calibration: A non-asymptotic theory and prediction-oriented optimality, 2018.

[3] Guillaume Damblin, Pierre Barbillon, Merlin Keller, Alberto Pasanisi, and Éric Parent. Adaptive numerical designs for the calibration of computer codes. *SIAM/ASA Journal on Uncertainty Quantification*, 6(1):151–179, Jan 2018.

[4] Valery V. Fedorov. Convex design theory 1. *Statistics*, 11:21–43, 1980.

[5] Paul Gardner, Charles Lord, and Robert Barthorpe. Sequential bayesian history matching for model calibration. 05 2019.

[6] Marc Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B*, 63:425–464, 02 2001.

[7] Luc Pronzato and Eric Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1):103–120, 1985.

[8] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.* Adaptive computation and machine learning. MIT Press, 2006.

Empirical study of the lengthscale parameters of anisotropic Matérn covariance in Gaussian processes with inactive variables

E. Bartok[†,1,2], N. Bonfils[§,2], M. Munoz Zuniga[§,2], E. Vazquez[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2022 – Nov. 2025

[1] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes (L2S)
{eniko.bartok,emmanuel.vazquez}@centralesupelec.fr

[2] IFP Énergies Nouvelles (IFPEN)
{nicolas.bonfils , miguel.munoz-zuniga}@ifpen.fr

Abstract

This study is conducted within the framework of wind turbine simulator modeling, which involves dealing with a multitude of parameters, some of which have minimal or negligible influence on the quantities of interest.

Gaussian Processes (GPs) [2, 5] are recognized for their effectiveness as metamodels [4]. They offer a Bayesian framework for supervised learning, allowing the incorporation of prior knowledge about a function through suitable kernel selection. The exploration of high-dimensional input in GPs has been investigated in various publications, with sensitivity analysis-based approaches, such as [1], representing just one avenue of exploration. Another approach is given in [6], where the authors use projections on subspaces.

In this work, we consider a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ with inactive variables, i.e. variables that do not influence the output. Specifically, our investigation focuses on the approximation accuracy and the lengthscale parameters of the Matérn covariance function [5], when they are selected by maximum likelihood estimation (MLE), in the presence of non-active variables.

More specifically, consider a constant-mean GP indexed by $\mathbb{R}^d$ with an anisotropic Matérn covariance $k_{\nu,\sigma,\rho}$ defined as

$$k_{\nu,\sigma,\rho}(x,y) := \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} h_\rho\right)^\nu K_\nu \left(\sqrt{2\nu} h_\rho\right),$$

where Γ is the Gamma function, K_ν is the modified Bessel function of the second kind,

$$h_\rho = \left[\sum_i^d \frac{(x_i - y_i)^2}{\rho_i^2}\right]^{1/2},$$

and $\nu \in \mathbb{R}^+$, $\sigma \in \mathbb{R}^+$ and $\rho = (\rho_1, \ldots, \rho_d) \in \mathbb{R}^{+d}$ are the parameters of the covariance. The ρ_i parameters are usually called lengthscale or range parameters.

Our contribution lies in the examination of the empirical distributions of lengthscale parameters associated with inactive variables.

In the community of GPs, it is well known that the lenghscale associated to inactive variables are typically large. [3] proves a theoretical result to support this: a lengthscale parameter ρ_i associated to a variable x_i tends towards infinity as the influence of x_i, measured as the norm of the partial derivative with respect to the variable, decreases.

In our study, we place ourselves in a non-asymptotic context, and we conduct an empirical study to provide practical insights. Specifically, we aim to determine the empirical distribution of lengthscale parameters for inactive variables as the design size varies when random maximin LHS are used for the design.

Moreover, our investigation explores how the number of inactive variables and the design size affect the distribution of lengthscale parameters, but also the convergence rate of the approximations.

We hope that this study will enable us to establish pruning procedures to reduce the dimension of our problems by looking at the lengthscale values ρ_i and identifying non-active variables.

Short biography (PhD student)

E. Bartok obtained her Master's degree in applied mathematics at Sorbonne Université in 2022. Her PhD project, which is funded by IFPEN, is entitled "Gaussian processes modeling for floating offshore wind turbine fatigue based on input/output dimension reduction". It focuses on the developpement of a new non-intrusive high-dimensional metamodeling strategy for the prediction of damage equivalent load (DEL) at a given point of the structure.

References

[1] Amandine Marrel, Bertrand Iooss, and Vincent Chabridon. The ICSCREAM methodology: Identification of penalizing configurations in computer experiments using screening and metamodel—applications in thermal hydraulics. *Nuclear Science and Engineering*, 196(3):301–321, 2022.

[2] Carl Edward Rasmussen, Christopher KI Williams, et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

[3] Malek Ben Salem, François Bachoc, Olivier Roustant, Fabrice Gamboa, and Lionel Tomaso. Sequential dimension reduction for learning features of expensive black-box functions. 2018.

[4] Thomas J Santner, Brian J Williams, William I Notz, Thomas J Santner, Brian J Williams, and William I Notz. Stochastic process models for describing computer simulator output. *The Design and Analysis of Computer Experiments*, pages 27–66, 2018.

[5] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

[6] Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando De Feitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

Improved High-Dimensional Covariance Matrix Estimation in Cross-Entropy Scheme for Rare Event Analysis

J. Beh[†,1,2], F. Simatos[§,2], J. Morio[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Jan. 2023 – Dec. 2025

[1] ONERA/DTIS, Université de Toulouse, F-31055 Toulouse, France
{jason.beh,jerome.morio}@onera.fr

[2] ISAE-SUPAERO and Université de Toulouse, Toulouse, France
florian.simatos@isae-supaero.fr

Abstract

In the context of the estimation of rare event probability $p_f(A) = \mathbb{P}_f(X \in A)$, where $f = N(0, I)$ is the d-dimensional standard Gaussian distribution (fairly general setting owing to isoprobabilistic transformations [5]), Importance Sampling (IS) estimator is often used due to the desired variance reduction property compared to Monte Carlo estimator. Given n_p samples Y_i generated according to g, an auxiliary distribution, to whom f is absolutely continuous, the IS estimator is written as

$$\hat{p}_f(A) = \frac{1}{n_p} \sum_{i=1}^{n_p} \ell(Y_i) \mathbb{1}\left(Y_i \in A\right),$$

where the importance weights $\ell(Y_i)$ denote $\frac{f(Y_i)}{g(Y_i)}$.

In high dimension, it is well known that IS estimators suffer from the weight degeneracy phenomenon: as $d \to \infty$,

$$\frac{\max_{i=1\dots n_p} \ell(Y_i)}{\sum_{i=1\dots n_p} \ell(Y_i)} \Rightarrow 1,$$

that is, the largest importance weight takes all the mass, hindering the convergence of the IS estimators. This phenomenon has been studied in our preprint with Y. Shadmi [2], from which it could be elicited that the main difficulty in high dimension is to ensure sufficient heaviness of the tail of g compared to that of f. In fact, it can be proven that if there exists $\alpha > 0$ such that $\sup_d \mathbb{E}_g(\ell(Y)^{1+\alpha}) < +\infty$, then $\max_{i=1\dots n_p} \ell(Y_i) / \sum_{i=1\dots n_p} \ell(Y_i) \Rightarrow 0$. Note that we always have $\mathbb{E}_g(\ell(Y)^1) = \mathbb{E}_f(1) = 1$, and the above implication is trivially true if we can take $\alpha = 1$ (the importance weights have finite variance).

When g is taken in the d-dimensional Gaussian family, say $g = N(\mu, \Sigma)$, with μ and Σ respectively the mean and the covariance matrix, $\ell(Y)$ involves the the following quadratic term in exponential, $\exp\left(-\frac{1}{2} Y^\top ((1+\alpha) I - \alpha \Sigma^{-1}) Y\right)$. To guarantee $\mathbb{E}_g(\ell(Y)^{1+\alpha}) < +\infty$, it is necessary that $\alpha < \alpha_*(\Sigma)$, where

$$\alpha_*(\Sigma) = \min\left(1, \frac{\lambda_1(\Sigma)}{1 - \lambda_1(\Sigma)}\right) = \begin{cases} \dfrac{\lambda_1(\Sigma)}{1 - \lambda_1(\Sigma)} & \text{if } \lambda_1(\Sigma) < \frac{1}{2}, \\ 1 & \text{else} \end{cases},$$

so that $(1+\alpha)I - \alpha\Sigma^{-1}$ is positive-definite. For instance, if the finite variance of the weights is desired ($\alpha = 1$), then $\lambda_1(\Sigma)$ has to be greater than $1/2$, a well-known result from [1]. This motivates the regulation of the behaviour of the smallest eigenvalue of Σ, which is in practice a random matrix estimated with adaptive importance sampling (AIS).

AIS consists in first estimating a suitable auxiliary density $\hat{g}$ with say n_g evaluations of the black-box function, then generating n_p samples from the learnt density in order to estimate $p_f(A)$ by IS. One popular scheme of AIS, Cross-Entropy (CE), attempts to estimate iteratively the optimal covariance matrix through empirical covariance matrix estimators. However, with a reasonable number of samples per iteration, the empirical estimators often underestimate the smallest eigenvalue of the optimal covariance matrix, cascading into weight degeneracy within the next few iterations of CE. We propose to adapt the Nonparametric Eigenvalue-Regularized COvariance Matrix Estimator (NERCOME) [4] into CE. NERCOME increases the smallest eigenvalue by decoupling the estimated eigenvalues and eigenvectors from the empirical estimators, therefore suitable for application into CE in order to avoid weight degeneracy. To further mitigate against the high dimensionality, NERCOME is projected onto its reduced eigenspace formed only by a select few eigenvectors, with their eigenvalues ranked increasingly according to $x \mapsto \log(x) - x + 1$, as suggested by recent results [3]. Some improvements on the probability estimation of several test cases up to dimension 300 have been obtained, which is then compared with state-of-the-art CE schemes [6, 7].

Short biography (PhD student)

I hold an M. Sc. in engineering from ISAE-SUPAERO in Toulouse. I completed my final-year internship at ISAE-SUPAERO which led to my current PhD thesis co-funded by EUR-MINT and ONERA, under the supervision of F. Simatos and J. Morio. The main goal of the thesis is to study the cause of the well-known curse-of-dimensionality for adaptive importance sampling, and to provide means to circumvent it based on this knowledge.

References

[1] S. Au and J. Beck. Important sampling in high dimensions. *Structural Safety* 25.2 (2003), pp. 139–163.

[2] J. Beh, Y. Shadmi, and F. Simatos. Insight from the Kullback–Leibler divergence into adaptive importance sampling schemes for rare event analysis in high dimension (2023). arXiv: `2309.16828 [math.ST]`.

[3] M. El Masri, J. Morio, and F. Simatos. Optimal projection to improve parametric importance sampling in high dimension (2022). arXiv: `2107.06091 [stat.CO]`.

[4] C. Lam. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *The Annals of Statistics* 44.3 (2016).

[5] R. Lebrun and A. Dutfoy. Do Rosenblatt and Nataf isoprobabilistic transformations really differ? *Probabilistic Engineering Mechanics* 24.4 (2009), pp. 577–584.

[6] I. Papaioannou, S. Geyer, and D. Straub. Improved cross entropy-based importance sampling with a flexible mixture model. *Reliability Engineering & System Safety* 191 (2019), p. 106564.

[7] F. Uribe et al. Cross-Entropy-Based Importance Sampling with Failure-Informed Dimension Reduction for Rare Event Simulation. *SIAM/ASA Journal on Uncertainty Quantification* 9.2 (2021), pp. 818–847.

Sensitivity analysis in numerical modelling-based morphodynamics

F. Ben Said [†,1,2], J. Reygner [§,2], K. El Kadi Abderrezzak [§,1,4], C. Goeury [§,1,4], A. Alfonsi [§,2], A. Dutfoy [§,3], F. Zaoui [§,1].

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Mar. 2023 – Mar. 2026

[1] EDF R&D, National Laboratory for Hydraulics and Environment (LNHE), Chatou, France
{faten.ben-said, kamal.el-kadi-abderrezzak, cedric.goeury, fabrice.zaoui}@edf.fr

[2] Center for Training and Research in MathematIcs and Scientific Computing (CERMICS) , Ecole des Ponts ParisTech, Champs-sur-Marne, France.
{julien.reygner, aurelien.alfonsi}@enpc.com

[3] EDF R&D, Performance and prediction of industrial risks for park simulation and studies (PERICLES), Paris-Saclay, France
anne.dutfoy@edf.fr

[4] Saint Venant Laboratory for Hydraulics (LHSV), Chatou, France

Abstract

Hydro-sedimentary modeling is affected from different sources of uncertainty which occur in process-based models, such as inaccuracies in the model inputs, errors in model structure, such as poorly described or omitted processes, and limited computing resources that constraints the number of parameters used to describe processes and the resolution of the model. This study aims to analyze and characterize these uncertainties, elucidating the factors contributing most significantly to the variability of model output. By employing Sobol sensitivity indices, a widely recognized variance-based sensitivity analysis approach, the aim of this work is to identify and mitigate influential factors, ultimately reducing model uncertainty.

In practice, Global sensitivity analysis based on Sobol's indices implying the stochastic estimation of statistical moments and indices is classically achieved with the Monte-Carlo technique. However, this technique requires a lot of computation time that can be reduced when a surrogate model is used in place of the hydraulic solver, for instance a polynomial chaos strategy [1,2]. Since non-linearity between processes in morphological models frequently result in gross feature of the model output being relatively insensitive to variations in inputs until some morphological threshold is reached, the accuracy of the Chaos Polynomial method can be undermined. Thus, this work introduces a solution to address this issue, presenting a novel and effective approach for simulating Sobol indices in the case of discontinuous models using the Chaos Polynomial method.

The hydro-informatic modeling system TELEMAC [3] is used to perform hydro-sedimentary numerical simulation on a channel bend. The sediment transport in a channel bend is subject not only to longitudinal transport, but also to transverse transport and transverse sorting by the secondary flow inherently associated with bends. The mechanics of sediment transport complexity is parametrized in modelling solver. After propagating the uncertainties inherent to the model, different methods to estimate the Sobol' sensitivity analysis are carried out. A

particular attention has been paid to proposing a relevant approach with Chaos Polynomial expansion on the discontinuous model.

The proposed method exhibits efficient performance compared to other techniques. The approach not only increase the precision of Sobol indices estimation but also provides a more cost-effective solution for uncertainty quantification in hydro-sedimentary modeling.

While the presented approach excels in scenarios with independent inputs, it poses challenges when dealing with dependent inputs. Future work aims to extend the methodology to address dependencies, further enhancing its applicability and robustness in real-world hydrodynamic modeling scenarios.

In conclusion, this study contributes to the field of uncertainty quantification in hydro-sedimentary modeling by introducing an innovative approach for simulating Sobol- sensitivity indices in the presence of discontinuous models. The proposed method demonstrates promising results, highlighting its potential to improve the accuracy and reliability of model predictions.

Figure 1: Illustration of the obtained results: a comparative analysis of classic and smart PC method applied to a hydro-sedimentology test case, focusing on the convergence of an input's Sobol indice with respect to the total degree of the meta-model.

Short biography (PhD student)

I'm a PhD student specializing in Applied Mathematics, specifically Probability and Statistics. Collaborating with the CERMICS Lab at Ecole des Ponts and EDF R&D LNHE department, my research centers on uncertainty quantification, characterization and taking into account of statistical dependencies in hydro-sedimentary modelling. This work is supported and funded by the French National Association of Research and Technology (ANRT) and EDF R&D with the Industrial Conventions for Training through Research (CIFRE grant agreement 2022/1501). My goal is to enhance the accuracy of hydro-sediment models, offering valuable insights for refining and advancing hydro-sediment modeling practices.

References

[1] B. Sudret, Global sensitivity using polynomial chaos expansion, 2006.

[2] G. Baltman, B. Sudret, An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis, 2008.

[3] J.M. Hervouet, Free Surface Flows, 2003.

Variational Bayes conditioning of a deep learning proxy

F. Bhavsar[†,1], N. Dessasis[§,1], F. Ors[§,1], T. Romary[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2021 – Nov. 2024

[1] Mines Paris, PSL University, Centre for geosciences and geoengineering, 77300 Fontainebleau, France
ferdinand.bhavsar@minesparis.psl.eu

Abstract

Stochastic process-based models have been used for decades for the modeling geological facies in the subsurface. However, their complexity poses a challenge when it comes to conditioning, i.e. fitting the model to field data. Recent developments in deep learning have led to an increased interest in neural networks as proxies for these models. Given sufficient quantity and quality of data, these deep generative algorithms match the realism of previous methods, but with more flexibility and faster generation.

In this paper we consider the problem of generating the spatial distribution of lithofacies in the subsurface and the subsequent task of generating simulations conditioned on field observations. This can be formulated as a Bayesian problem where we first learn a prior distribution $p(x)$, and then the posterior distribution $p(x|x^\star)$, where x is the lithofacies realization and $x^\star$ is the field data. We focus on Flumy [4], a stochastic process-based modeling tool for meandering channels in both fluvial and turbidite environments. We use this software to generate non-conditional realisations, which are used as a dataset to train a faster and more flexible deep learning proxy.

To learn the prior model, we use the Generative Adversarial Networks (GAN) framework [3], which pits a generator neural network against a discriminator neural network to find a Nash equilibrium. They exhibit state-of-the-art results when generating fake realizations of multivariate distributions such as images [3, 1]. The generator produces synthetic data x by sampling from a Gaussian prior distribution and transforming it through a function $G(z, \theta^G)$, where z is a realization of the input distribution and θ^G are the parameters of the generator. Through G, the model thus defines the synthetic data distribution p_θ. The discriminator assigns, to each data point it receives as input, a probability that this data point was sampled from the training data distribution $p_r(x)$. Formally, it defines a family of functions $D(x, \theta^D)$ from the realization space to the interval $[0, 1]$, where x is a data point and θ^D are the parameters of the discriminator. Both models are trained alternatively, as a zero-sum game, using stochastic gradient descent.

Once the prior distribution is approximated using the generator model, variational Bayesian methods can be used to approximate a posterior distribution, $p(x|x^\star)$. When generating realizations through the generator, $x = G(z)$, where z is the input to the generator, we can instead consider $p(z|x^\star)$. This posterior distribution is estimated by learning the parameters Ψ of an inference model, that defines the parametric distribution q_Ψ. By using variational Bayesian methods, we can use common optimization methods to estimate the optimal parameters Ψ^* minimizing the Kullback-Leibler divergence between the real distribution $p(z|x^\star)$ and $q_\Psi(z|x^\star)$:

$$q^*_\Psi(z|x^\star) = \underset{q_\Psi(z|x^\star)\in\mathcal{Q}}{\arg\min}\, D_{KL}\Big(q_\Psi(z|x^\star) \;||\; p(z|x^\star)\Big) \tag{1}$$

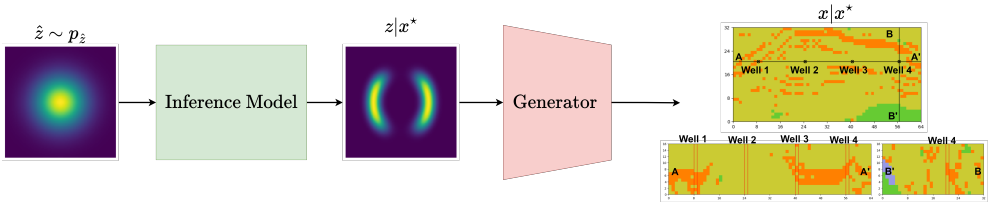


Figure 1: Diagram of the conditional generation workflow where we use a secondary model that transforms a random vector $\hat{z}$ into the conditioned variable $z|x^\star$.

In previous literature, this inference model is a neural network [2]. However, we present here our method where the inference model is instead a Gaussian Mixture instead. This new method improves the variety of conditioned simulations that can be generated.

In our method, we choose a number of Gaussian components K. Each component $i = 1, ..., K$ is defined by its mean μ_i and its covariance matrix Σ_i, as well as a non-negative weight π_i. The weights π_i sum to 1. The parametric distribution is then defined as $q_\Psi(z) = \sum_{i=1}^{K} \pi_i f_{\mu_i,\Sigma_i}(z)$, where $f_{\mu,\Sigma}$ is the Gaussian density with mean μ and covariance Σ.

The proposed approach is tested on 2D and 3D simulations. The GAN model presented uses several stabilisation techniques from recent literature, such as W-GANs, multiscale approaches and spectral normalization. The non-conditional results are evaluated through morphological properties (e.g., facies connectivity and transition probability) and the conditional results by accuracy and variability metrics. We discuss the results as well as shortcomings and perspectives of our approach.

Short biography (PhD student)

Ferdinand Bhavsar is a third-year PhD student in the Geostatistics team at Mines Paris, PSL University, Centre for geosciences and geoengineering. He graduated from EPITA in 2021 with a focus on deep learning methods. His thesis subject is the development of geostatistical spatio-temporal methods of simulation using generative models of deep learning. This thesis is founded by the chair Carnot M.I.N.E.S Industrie Responsable.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis, 2019.

[2] Shing Chan and Ahmed H. Elsheikh. Parametric generation of conditional geological realizations using generative neural networks. *Computational Geosciences*, 23(5):925–952, jul 2019.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] Simon Lopez. Channelized reservoir modeling: a stochastic process-based approach. 06 2003.

Coupled Parameter-Data Dimension Reduction for Bayesian Inverse Problems: Application to Optimal Sensor Placement

Q. Chen[†,1], E. Arnaud[§,1], R. Baptista[2], O. Zahm[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2021 – Sep. 2024

[1] LJK, Université Grenoble Alpes, AIRSEA Inria Grenoble
{qiao.chen,elise.arnaud,olivier.zahm}@inria.fr

[2] CMS, California Institute of Technology
rsb@caltech.edu

Abstract

Dimension reduction is crucial for Bayesian inverse problems involving numerous simulations with various parameters. Considerable research has focused on reducing the high-dimensional parameter space resulting from discretization of fields [2]. Recently, interest has grown in reducing the data space to compute optimal experimental designs for Bayesian inverse problems. However, the reduction of both spaces has traditionally been treated separately, even though choosing a low-dimensional parameter subspace influences which data subspace is informative and vice versa.

We propose a method that couples the dimension reduction of both spaces, and controls the posterior approximation error by exploiting access to gradients. Building on the work in [4, 1], we reduce the parameter $X \in \mathbb{R}^d$ and data $Y \in \mathbb{R}^m$ by retain only some projected components

$$\begin{aligned} X_r &= U_r^\top X, && \text{where} \quad U_r \in \mathbb{R}^{d\times r}, \\ Y_s &= V_s^\top Y, && \text{where} \quad V_s \in \mathbb{R}^{m\times s}. \end{aligned}$$

In Bayesian inference, the objective is typically to find U_r and V_s minimizing the posterior approximation error in the expected Kullback-Leibler divergence $\mathbb{E}_Y[\mathbb{D}_{\mathrm{KL}}(\pi_{X|Y}\|\widetilde{\pi}_{X|Y})$. This can be traced back to minimising the L^2-error $\mathbb{E}[\|G(X) - \widetilde{G}(X)\|^2]$ of the forward operator $G : \mathbb{R}^d \to \mathbb{R}^m$, where $Y = G(X) + \varepsilon$ with Gaussian noise ε. We prove that the optimal U_r and V_s for a Poincaré-type upper bound of the L^2-error consist of the dominant eigenvectors of the diagnostic matrices

$$\begin{aligned} H_X(V_s) &= \mathbb{E}\left[\nabla G(X)^\top V_s V_s^\top \nabla G(X)\right], \\ H_Y(U_r) &= \mathbb{E}\left[\nabla G(X) U_r U_r^\top \nabla G(X)^\top\right] \end{aligned}$$

respectively. The projectors can be computed efficiently by doing an alternating singular value decomposition on the diagnostic matrices. Additionally, our method allows for goal-oriented dimension reduction by fixing some parameter of interest $X_r = U_r^\top X$ and computing the corresponding V_s that best captures the resulting output variations.

As potential application, we investigate Bayesian optimal experimental design problems, which aim to strategically place sensors to minimize uncertainty in the estimated parameters. These

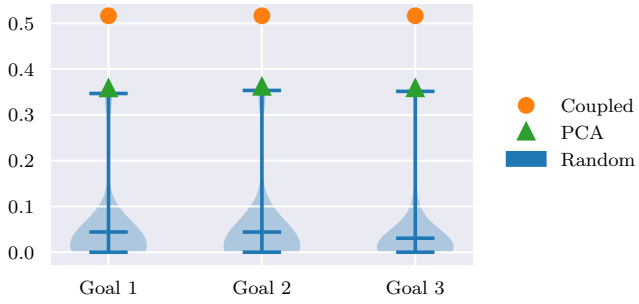


Figure 1: Expected information gain of sensor placement computed using coupled dimension reduction compared to PCA-based and random sensor selection.

problems present an even greater challenge than Bayesian inference as they involve the inverse problem as a sub-problem. We show how our goal-oriented data dimension reduction objective relates to maximizing the expected information gain of the experimental design V_s

$$\max_{V_s} \mathbb{E}_Y \left[\mathbb{D}_{\mathrm{KL}} \left(\pi_{X_r|Y_s} \middle\| \pi_{X_r} \right) \right]$$

for some parameter of interest X_r. Existing sensor placement methods are often limited to Gaussian posteriors due to the costly evaluation of the expected information gain. Furthermore, we avoid combinatorial optimization over all possible sensor selections by applying the fast empirical interpolation method [3] on our optimal subspace V_s.

Short biography (PhD student)

I graduated with a MSc in mathematics from Heidelberg University, Germany and am currently a third-year PhD student in the inria team AIRSEA at the Université Grenoble Alpes. My thesis aims to develop dimension reduction methods for Bayesian inverse problems that are applicable for ocean and atmosphere science. It is funded by the Université Grenoble Alpes.

References

[1] Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. Gradient-based data and parameter dimension reduction for bayesian models: an information theoretic perspective. *arXiv preprint arXiv:2207.08670*, 2022.

[2] Paul G Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014.

[3] Yvon Maday, Ngoc Cuong Nguyen, Anthony T Patera, and George SH Pau. A general, multipurpose interpolation procedure: the magic points. 2007.

[4] Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, and Youssef Marzouk. Certified dimension reduction in nonlinear bayesian inverse problems. *Mathematics of Computation*, 91(336):1789–1835, 2022.

Identification of the excitation of trains from on-board measures

M. Chihaoui[†,1], D. Duhamel[§,1], G. Perrin[§,2], C. Funfschilling[§,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Mar. 2023 – Mar. 2026
malek.chihaoui@sncf.fr

[1] Laboratoire Navier, Ecole Nationale des Ponts et Chaussées
denis.duhamel@enpc.fr

[2] COSYS, Université Gustave Eiffel
guillaume.perrin@univ-eiffel.fr

[3] SNCF, Direction Technologies, Innovation et Projets Groupe
christine.funfschilling@sncf.fr

Abstract

The knowledge of the axle-rail contact forces and the tracks irregularities is essential to study both the *dynamic response* and the *mechanical damage* of the train and the tracks. Such information would be an input to many computational models (Rail and Wheel damage, Track settlement, Wheel wear...) and would considerably enhance the optimization of future conception and maintenance. Nevertheless, these are relatively poorly known due to the difficulty in measuring them on commercial trains. Nowadays, special trains which regularly monitor the national rail network are used to measure the track irregularities for maintenance. On the other hand, the wheel/rail loads are measured only during special on-track test campaigns with measuring wheelsets. All these trains are not authorized to take regular passengers. Their usage is therefore costly and limited.

The present work focuses on the reconstruction of the rails geometry and the contact forces between the train axles and the tracks with the help of on-board sensors which can be used on commercial trains. For the moment we work on a numerical experiment: the dynamic response of the train is estimated thanks to a railway dynamic simulation which takes as input measured tracks irregularities. The inverse problem as shown in the figure 1 is however complex to solve [1, 2].

That is why we first worked on the understanding of the mechanical impact that track irregularities $\boldsymbol{d}$ have on the kinematics of the train $\boldsymbol{Y}$ using the same numerical simulation software. This allows us to construct a robust mathematical estimator of the track irregularity based on the measured dynamic response. Having a good enough estimate of the excitation for each train response, the next step is to refine it to find the excitation imposed on the train that would produce a similar response. This makes our problem an optimization problem.

The first difficulty we encounter lies in the proper mathematical definition of the identification of this excitation which must be robust to the *uncertainties* to which the train is subject (Wind, Humidity, Load...) but also to the measure, model & numerical errors [3, 4]. Secondly, it must be said that the track-train system is highly non-linear which constitutes a non-negligible

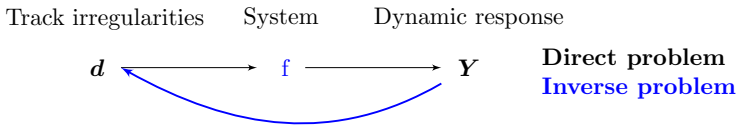


Figure 1: Direct & inverse problems diagram

difficulties. Thirdly, for the inverse problem, a functional excitation must be found for the optimization problem. For this, we intend to use the Covariance Matrix Adaptation - Evolution Strategy (CMA-ES) algorithm. *Dimensionnality reduction* and use of *meta-models* must also be considered to achieve the optimization.

References

[1] D. Lebel, C. Soize, C. Funfschilling, and G. Perrin, "Statistical inverse identification for nonlinear train dynamics using a surrogate model in a bayesian framework," *Journal of Sound and Vibration*, pp. 158–176, 2019.

[2] D. Lebel, C. Soize, C. Funfschilling, and G. Perrin, "High-speed train suspension health monitoring using computational dynamics and acceleration measurements," *Vehicle Systems Dynamics*, pp. 911–932, 2020.

[3] G. Schueller and H. J. Pradlwarteren, "Uncertain linear systems in dynamics : Retrospective and recent developments by stochastic approaches," *Engineering Structures*, pp. 2507–2517, 2009.

[4] A. Nouy, "Proper generalized decomposition and separated representations for the numerical solution of high dimensional stochastic problems," *Archives of Computational Methods in Engineering*, pp. 4030–4034, 2010.

Short biography (PhD student)

Malek Chihaoui is currently an PhD student at l'École Nationale des Ponts et Chaussées. He is supervised by Prof Denis Duhamel, Prof Guillaume Perrin and Dr Christine Funfschilling. His thesis focuses on the reconstuction of the irregularities of the tracks and the axle-rails contact forces with the help of on-board sensors on the train. The work is supported by the French railway company SNCF.

Surface area and volume on point cloud based polytopic tessellations

R. Cotsakis[†,1], E. Di Bernardino[§,1], C. Duval[2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: May 2021 – May 2024

[1] Université Côte d'Azur, Laboratoire J.A. Dieudonné, UMR CNRS 7351, Nice, France
[†] Ryan.Cotsakis@unice.fr

[2] Université de Lille, Laboratoire Paul Painlevé, UMR 8524 CNRS, Lille, France

Abstract

The excursion set of a C^2 smooth random field carries relevant information in its various geometric measures. From a computational viewpoint, one never has access to the continuous observation of the excursion set, but rather to observations at discrete points in space. It has been reported that for specific regular lattices of points in dimensions 2 and 3, the usual approximation of the surface area of the excursions does not converge when the lattice becomes dense in the domain of observation to the desired limit. In the present work, under the key assumptions of stationarity and isotropy, we demonstrate that this limiting factor is invariant to the locations of the observation points. Indeed, we identify an explicit formula for the correction factor, showing that it only depends on the spatial dimension d. This enables us to define an approximation for the surface area of excursion sets for general tessellations of polytopes in $\mathbb{R}^d$, including Poisson-Voronoi tessellations. We also establish a joint central limit theorem for the surface area and volume of excursion sets observed over hypercubic lattices.

Short biography (PhD student)

Ryan specializes in computational geometry, stochastic geometry, and extreme value theory.

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. This work has been partially supported by the project ANR MISTIC (ANR-19-CE40-0005).

Bayesian calibration of innovative sensors for monitoring pollutants in uncontrolled environments

Marine Dumon[†,1], Bérengère Lebental[§,1], Guillaume Perrin[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2022 – Sep. 2025

[1] COSYS/IMSE, Université Gustave Eiffel
marine.dumon@univ-eiffel.fr

Abstract

Air and water pollution are now a major public health issue, and developing efficient detection methods to better monitor this pollution is essential to reduce the risk of exposure. Innovative materials such as nanomaterial-based sensors [3] have been proposed for their high sensitivity to different chemical species in air and water and their ability to detect them even in very low concentrations. However, if in laboratory these sensors are able to show encouraging results, the passage in real conditions can pose difficulties. Indeed, current calibration methods may not always yield precise measurements of air and water quality when inexpensive or innovative instruments are deployed in real environments, as they are disturbed by temperature or relative humidity and often encounter unmeasured interferents (highly sensitive but not very selective). Moreover for these sensors, there are no physical models, meaning that the relationship between these variables is entirely unknown. Accurately modelling this relationship, including sensor outputs, target quantities, and interfering quantities, in a non linear manner is a difficult task.

Two distinct phases are therefore needed for their use in uncontrolled environments. In the first step, the dependence of sensor responses on pollutant concentrations is estimated using labelled data provided by reference sensors. In the second step, this relationship is used to predict the pollutant concentration from the sensor outputs only. This estimation is made challenging by the presence of all the sources of uncertainty: the potential existence of unmeasured but influential pollutants, the measurement noises for the input and output data, and the fact that the relationship between sensors inputs and outputs is only partially learned. In order to integrate all these sources of uncertainty and make robust predictions, we propose to solve the problem using a Bayesian formalism [2].

The key point of this formalism is the choice of inputs and outputs. The difficulty is to choose the real influential parameters. A sensitivity analysis [1] must be conducted to find the parameters, among all the possible chemical species measured, having a causal impact on the sensor outputs. However, these pollutants are highly correlated and interact with each other through a relationship that is not trivially known. Conditional independence allows us to mathematically understand the real dependencies between sensor outputs and pollutants. We can then improve the calibration model by limiting the number of input variables.

Short biography (PhD student)

With a mathematical background, Marine Dumon is currently a PhD student at Gustave Eiffel University under the direction of Bérengère Lebental and Guillaume Perrin. Her thesis focuses

on innovative calibration methods for air and water quality sensors and is funded by innovative sensor design and deployment projects.

References

[1] Margot Herin, Marouane Il Idrissi, Vincent Chabridon, and Bertrand Iooss. Proportional marginal effects for global sensitivity analysis. 2022.

[2] Bérengère Lebental Guillaume Perin. Uncertainty-based calibration method for environmental sensors - application to chlorine and ph monitoring with carbon nanotube sensor array. *IEEE sensors journal*, 2023.

[3] Deepika Tyagi Huide Wang Weichun Huang Lanping Hu Yanfeng Tang Zhinan Guo Zhengbiao Ouyang Han Zhang. Recent advances in two-dimensional-material-based sensing technology toward health and environmental monitoring applications. *Nanoscale*, 12:3535–3559, 2005.

Easy conditioning way beyond the Gaussian case.

A. Faul[†,1], D. Ginsbourger[§,1], B. Spycher[§,2],

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2022 – Sep. 2026

[1] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland
antoine.faul@unibe.ch david.ginsbourger@unibe.ch

[2] Institute of Social and Preventive Medicine, University of Bern, Switzerland
ben.spycher@unibe.ch

Abstract

In many statistical and machine learning settings, one is interested in estimating the conditional distribution of a random vector $\mathbf{Y} \in \mathbb{R}^q$ given an observation of $\mathbf{X} \in \mathbb{R}^p$, based on a sample of $(\mathbf{X}, \mathbf{Y})$ previously acquired. Assuming for instance $(\mathbf{X}, \mathbf{Y})$ to possess a probability density function $f(\mathbf{x}, \mathbf{y})$ such that $f(\mathbf{x}) = \int_{\mathbb{R}^q} f(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y} \neq 0$, the target can be characterized in terms of the conditional density function $f(\mathbf{y}|\mathbf{x}) = f(\mathbf{x}, \mathbf{y})/f(\mathbf{x})$, where the notation f is colloquially used to speak of the joint, marginal, and conditional densities (distinguished by their arguments). Now, given that f is generally not known and rather estimated from the considered sample, resulting in an estimate joint density $\hat{f}$, a natural approach to estimate the conditional density $f(\mathbf{y}|\mathbf{x})$ is to take $\hat{f}(\mathbf{y}|\mathbf{x}) = \hat{f}(\mathbf{x}, \mathbf{y})/\hat{f}(\mathbf{x})$. In some cases, conditional distributions can be derived in closed form and/or are easy to sample from. This is notably the case for multivariate Gaussian, Student [2], and Skew-Normal distributions [1]. However, assuming more general forms for $\hat{f}$, conditioning can become less straightforward and call for potentially cumbersome approximate numerical procedures such as Markov chain Monte Carlo methods or variational inference. Our focus here is on extending the easy conditioning properties of the aforementioned distribution classes to broader classes obtained by mixtures and transformations, thus achieving more expressiveness while still enjoying efficient conditioning.

The first property underlying our main approach concerns the propagation of easy conditioning from components to resulting mixtures. We focus here on Borel sets B of $\mathbb{R}^q$ and on elements $\mathbf{x} \in \mathbb{R}^p$, with aim to calculate $\mathbb{P}(\mathbf{Y} \in B|\mathbf{X} = \mathbf{x})$ (resp. $f(\mathbf{y}|\mathbf{x})$) efficiently in cases where the joint distribution of $(\mathbf{X}, \mathbf{Y})$ is assumed to a finite mixture of distributions under each of which $\mathbf{Y}$ is easy to condition on $\mathbf{X} = \mathbf{x}$. If there exists indeed $K \in \{1, 2, \dots\}$ and a latent Z (defined on the same probability space as $(\mathbf{X}, \mathbf{Y})$) taking values in $\{1, 2, \dots, K\}$ such that, for any $i \in \{1, 2, \dots, K\}$ and $\mathbf{x} \in \mathbb{R}^p$, $\mathbb{P}(\mathbf{Y} \in B|\mathbf{X} = \mathbf{x}, Z = j)$ and $\mathbb{P}(Z = j|\mathbf{X} = \mathbf{x})$ are easily calculated, then we enjoy a facilitated calculation of the target probability by noticing that,

$$\mathbb{P}(\mathbf{Y} \in B|\mathbf{X} = \mathbf{x}) = \sum_{j=1}^{K} \mathbb{P}(\mathbf{Y} \in B|\mathbf{X} = \mathbf{x}, Z = j)\mathbb{P}(Z = j|\mathbf{X} = \mathbf{x})$$

Now, when $(\mathbf{X}, \mathbf{Y})$ admits a density $f = \sum_{j=1}^{K} \alpha_j f_j$ with $\alpha_1, \dots, \alpha_K > 0, \sum_{j=1}^{K} \alpha_j = 1$, the conditional density can similarly be expressed as mixture of conditional densities via

$$f(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{K} \widetilde{\alpha}_j(\mathbf{x}) f_j(\mathbf{y}|\mathbf{x})$$

with $\widetilde{\alpha}_j(\mathbf{x}) = \alpha_j f_j(\mathbf{x})/f(\mathbf{x})$, where $f(\mathbf{x}) \neq 0$ is assumed. As a consequence, if the $f_j(\mathbf{y}|\mathbf{x})$'s and $f_j(\mathbf{x})$'s are known in closed form, so is $f(\mathbf{y}|\mathbf{x})$. Obviously, one also enjoys such a fast conditioning when $\hat{f}$ is assumed to be in a relevant class, i.e. a finite mixture of "easy to condition" joint distributions. This encompasses in the first place mixtures of multivariate Gaussians, but also mixtures of multivariate Student t-distributions (possibly with different location and scatter parameters for the different components), as well as mixtures of Skew-Normals.

Another property relates to the extension of the "easy to condition" properties to block-wise transformations. Let consider that, $\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} = \begin{bmatrix}\phi_1(\mathbf{U})\\ \phi_2(\mathbf{V})\end{bmatrix}$ where ϕ_1 and ϕ_2 are two bijections, and $\mathbf{V}$ is easy to condition on $\mathbf{U} = \mathbf{u}$. Then with the same notations as above, we obtain:

$$\mathbb{P}(\mathbf{Y} \in B|\mathbf{X} = x) = \mathbb{P}(\mathbf{V} \in \phi_2^{-1}(B)|\mathbf{U} = \phi_1^{-1}(\mathbf{x}))$$

Thus, we can calculate easily the target probability. Furthermore, when we only assume that the joint distribution of $(\mathbf{U}, \mathbf{V})$ is a finite mixture of distributions under each of which $\mathbf{V}$ is easy to condition on $\mathbf{U} = \mathbf{u}$, and such that, $\forall i \in \{1, 2, \ldots, K\}$ and $\mathbf{u} \in \mathbb{R}^p$, $\mathbb{P}(\mathbf{V} \in B|\mathbf{U} = \mathbf{u}, Z = j)$ and $\mathbb{P}(Z = j|\mathbf{U} = \mathbf{u})$ are easily calculated, then by combining our two properties we can readily compute the target probability and/or sample from the conditional distribution.

By considering marginal probability transforms, our generic results tell us in particular that if we assume the dependence structure of the joint distribution to follow a mixture copula with components "easy to condition", such as the Gaussian mixture copula [3] or the Student t mixture copula , then we still enjoy an efficient calculation of the target conditional probability and we are able to perform fast conditional sampling.

Our current research encompasses the use of this method for probabilistic imputation of missing values and for conditional density estimation. We will demonstrate the versatility and the computational benefits of our approach on real and artificial data sets motivated by problems from various fields such as medicine and astrophysics.

Short biography (PhD student)

Antoine Faul graduated with a MSc in Engineering from Isae-Supaero and with a MSc in Statistics from Université Paris-Saclay. He started his PhD at the University of Bern in October 2022 on statistics with application to medicine. The thesis is funded by the Multidisciplinary Center for Infectious Diseases (MCID) of the University of Bern.

References

[1] Adelchi Azzalini and A Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

[2] Peng Ding. On the conditional distribution of the multivariate t distribution. *The American Statistician*, 70(3):293–295, 2016.

[3] Ashutosh Tewari, Michael J Giering, and Arvind Raghunathan. Parametric characterization of multimodal distributions with non-gaussian modes. In *2011 IEEE 11th international conference on data mining workshops*, pages 286–292. IEEE, 2011.

Hydrologic model calibration using Magnetic Resonance Sounding (MRS) measurements

G. Gru[†,1], N. Lesparre[§,1] P. Ackerer[§,1], B. Belfort[c,1], J.F. Girard[c,1]

[†] PhD student (presenting author). [§] PhD supervisors [c,1] Collaborators

PhD expected duration: Nov. 2023 – Oct. 2026

[1] Institut Terre et Environnement de Strasbourg, Université de Strasbourg/EOST/ENGEES, CNRS UMR 7063, 5 rue Descartes, Strasbourg F-67084, France
{gru,lesparre,ackerer}@unistra.fr

Abstract

Water ressources in montainous areas are of a primordial importance both for local ecosystems and for human activities [3]. Prediction uncertainties concerning the evolution of the water ressource have to be quantified and reduced. The aim of this work is to use geophysical sounding data to monitor the water stock in the subsurface and its evolution. This information is then used to estimate the parameters of a hydrologic model using automatic calibration methods. We have coupled a 1D hydrologic model based on Richards equation with a geophysical model. The geophysical model is used to compute Magnetic Resonance Sounding (MRS) signals from the output of the hydrologic model: water content distributions. This coupled model is used to estimate the hydrodynamic parameters with Bayesian inference methods. The one dimensional water flow model we are working with is based on a PDE : Richards equation. We are using a numerical code named WAMOS (WAter MOvement in Soil), that solves the PDE using a Newton-Raphson method [1]. The geophysical monitoring technique used is MRS, and its peculiarity is that it is directly sensitive to the water content. The idea of this method is similar to magnetic resonance imaging in the health science field. An electomagnetic field is created, energizing hydrogen nuclei contained in water molecules in the subsurface. Then the response of the subsurface is recorded and a water content distribution is deduced from the surface signal [2].

The experimental site where the data set was collected is the Strengbach catchment. It is an experimental water catchment located in the Vosges mountains, France (`ohge.unistra.fr`). This experimental site is part of the Critical Zone observatories network (`www.ozcar-ri.org`). Meteorologic, hydrologic and geochimical data have been acquired on this experimental site since 1986.

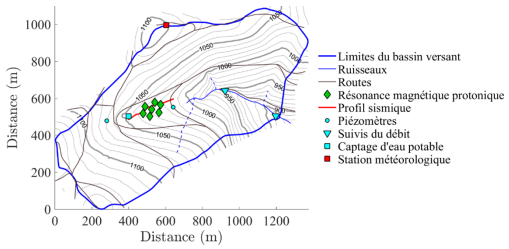


Figure 1: Map of the Strengbach catchment with the measurement stations.

Meteorologic and piezometric data are used along with an evapotranspiration model to constrain the boundary conditions of the hydrologic simulation. The upper (flux) boundary condition is constrained by the measured precipitation processed by an evapotranspiration model and a snow model. The lower (pressure) boundary condition is constrained by the piezometric level acquired at a piezometer close to the MRS station. Data acquisition frequency is 1 measure every 10 minutes.

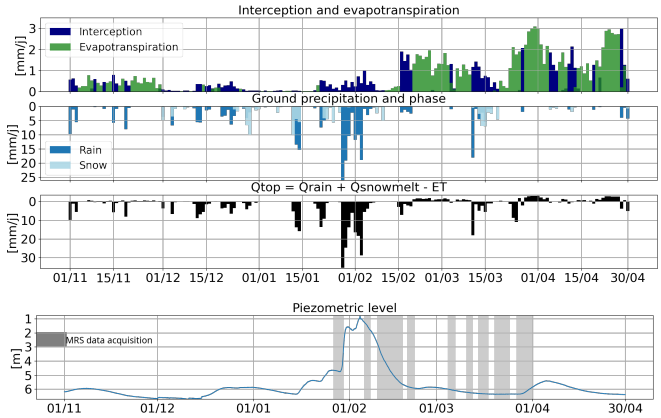


Figure 2: Meteorologic and hydrologic data used to constrain boundary conditions (year 2020-2021).

Short biography (PhD student)

I graduated from a bachelor of applied mathematics at the university of Rennes 1. Then I pursued with a Masters of applied mathematics in the same university. The VHYPERE (Valeur Ajoutée de l'Hydrogéophysique pour la Prédiction de la Ressource en Eau en montagne) project aims to use and quantify the value of geophysical data to study the hydrodynamic properties of the soil in montainous water catchments. This PhD was funded by a grant from the doctoral school of earth and environmental sciences of Strasbourg (ED413).

References

[1] F. Hassane Maina and P. Ackerer. Ross scheme, newton–raphson iterative methods and time-stepping strategies for solving the mixed form of richards' equation. *Hydrology and Earth System Sciences*, 21(6):2667–2683, 2017.

[2] A. Legchenko and P. Valla. A review of the basic principles for proton magnetic resonance sounding measurements. *Journal of Applied Geophysics*, 50(1-2):3–19, 2002.

[3] D. Viviroli et al. Climate change and mountain water resources: overview and recommendations for research, management and policy. *Hydrology and Earth System Sciences*, 15(2):471–504, 2011.

Weighted Poincaré inequalities in dimension 1 and global sensitivity analysis

D. Heredia[†,1], A. Joulin[§,2], O. Roustant[§,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2023 – Sep. 2026

[1,2,3] INSA Toulouse, Institut de Mathématiques de Toulouse
david.heredia@insa-toulouse.fr
alderic.joulin@insa-toulouse.fr
olivier.roustant@insa-toulouse.fr

Abstract for a poster at Mascot-Num 2024

Functional inequalities and global sensitivity analysis (GSA) are two highly active and seemingly distinct mathematical domains. The field of functional inequalities, that is part of the functional analysis domain and whose foundations date back to at least the early 20th century with the work of Henri Poincaré, studies general inequalities that hold for an entire class of functions, typically in a Sobolev space. The subfield of statistics GSA is more recent and falls within an explicability framework. It seeks to quantify the influence of input variables on the output of a multivariate function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, which is expensive to evaluate [1]. These variables can represent calculation codes that model complex phenomena or artificial intelligence algorithms whose functioning is not well understood.

An important connection between the two domains was initiated in 2013 by [3], where it is shown that the inequality that arises between two sensitivity indices (Sobol and DGSM indices) directly stems from a Poincaré inequality. This leads to address Poincaré inequalities from the perspective of the explicability of calculation codes and artificial intelligence algorithms.

Classical Poincaré inequalities have been extensively revisited in dimension 1 from a GSA perspective (see for example [4]). As a new proposal for the application of functional inequalities to GSA, we develop the use of *weighted Poincaré inequalities* in dimension 1. These are similar to the classical ones but include a non negative weight introduced on the right-hand side of the inequality. The use of weights provides an additional degree of freedom that can be manipulated to enhance the precision of the inequalities, and it is sometimes necessary for certain probability distributions that do not satisfy a classical Poincaré inequality (e.g., the Cauchy distribution).

In GSA, [5] and independently [1] suggest, by using the dual spectral interpretation of these weighted Poincaré inequalities, to choose the weight in such a way that the first eigenfunction of the diffusion operator associated with the inequality is linear, with a perspective of linearizating f. We extend their approach, which was limited to a few examples of probability distributions, with the implementation of a numerical method that approximates the weight associated with any probability measure and any function that is forced to be the first eigenfunction (not only the linear one). We also provide a theoretical justification by considering the dual spectral problem associated with the inequality.

[1] There is, of course, a more general setting that the one described here, and GSA is not limited to the study of functions defined from $\mathbb{R}^d$ to $\mathbb{R}$.

We test the effectiveness of the corresponding weighted Poincaré inequalities with a simplified flood model that has been addressed in various works for pedagogical reasons (see, for example, [2] and [4]).

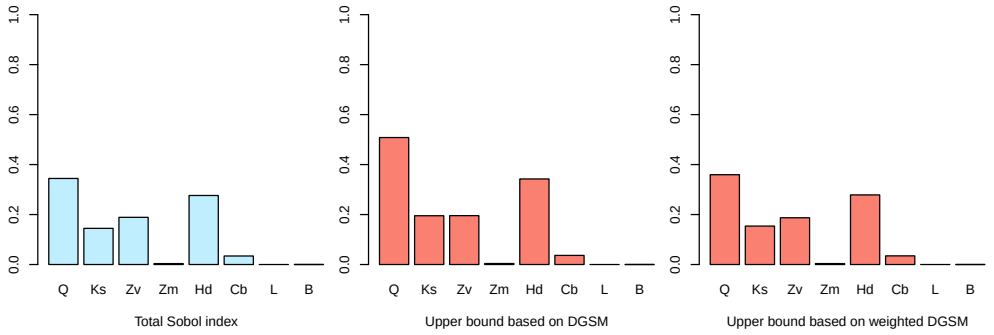


Figure 1: Barplots containing the Total Sobol indices (left one) and their upper bounds based on the DGSM indices (middle one) and the *weighted* DGSM indices (right one), of the maximal annual overflow of a river, that depends on random variables Q, K_s, Z_v, Z_m, H_d, C_b, L and B.

Both DGSM indices and weighted DGSM indices are used to upper bound the Total Sobol Indices, which are computationally expensive to estimate. We can observe the improvement of the upper bounds when we use a weight. The improvement is notable for the case of the random variable Q, which follows a truncated Gumbel distribution, distribution that naturally appears in extreme value theory and that is classically employed in hydrology.

Short biography (PhD student)

After completing my mathematics studies in Ecuador, my home country, the Mathematical Foundation Jacques Hadamard (FMJH) provided me with the oportunity to continue my formation at the prestigious Paris-Saclay University. Upon completing my master's degree last year, I secured a ministerial doctoral contract at INSA Toulouse and at the Toulouse Mathematics Institute (IMT), under the supervision of Aldéric Joulin and Olivier Roustant.

References

[1] Gilles Germain and Yvik Swan. A note on one-dimensional poincaré inequalities by stein-type integration, 2022. arXiv:2112.02638.

[2] Bertrand Iooss and Paul Lemaître. *A Review on Global Sensitivity Analysis Methods*, pages 101–122. Springer US, Boston, MA, 2015.

[3] Matieyendou Lamboni, Bertrand Iooss, Anne-Laure Popelin, and Fabrice Gamboa. Derivative-based global sensitivity measures: general links with Sobol' indices and numerical tests. *Mathematics and Computers in Simulation*, 87:45–54, 2013.

[4] Olivier Roustant, Franck Barthe, and Bertrand Iooss. Poincaré inequalities on intervals – application to sensitivity analysis. *Electronic Journal of Statistics*, 11(2):3081 – 3119, 2017.

[5] Shufang Song, Zhou Tong, Lu Wang, Sergei Kucherenko, and Zhenzhou Lu. Derivative-based new upper bound of sobol' sensitivity measure. *Reliability Engineering & System Safety*, 187:142–148, 05 2018.

An adaptive approach for surrogate modeling based on Full Maximum a Posteriori Estimation of model error for efficient calibration of numerical codes

S. Janati Idrissi[†,1,3], P.M. Congedo[§,1], O. Le-Maître[§,2], M.G. Rodio[§,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2022 – Sep. 2025

[1] Inria, Centre de Mathématiques appliquées, Ecole Polytechnique, IPP, Route de Saclay, 91128 Palaiseau, France
{olivier.le-maitre, pietro.congedo}@polytechnique.edu

[2] CNRS, Inria, Centre de Mathématiques appliquées, Ecole Polytechnique, IPP, Route de Saclay, 91128 Palaiseau

[3] Commissariat à l'Energie Atomique et aux Energies alternatives, ISAS, D36, 91191 Gif-sur-Yvette, France
{sanae.janatiidrissi, mariagiovanna.rodio}@cea.fr

Abstract

Computer codes are essential in nuclear engineering. They are used, for instance, to understand reactor kinetics and thermal-hydraulic processes and to study extreme scenarios and accidental conditions. However, these codes are based on complex models involving several context-specific parameters that may be unknown. In order to infer the best values of these parameters, model predictions are compared to experimental observations. Besides, all these models are imperfect and can not reproduce real-world processes. The difference between model predictions and reality is due to the model error, and it should be accounted for in the identification of the model parameters [1].

In this work, we rely on a Bayesian calibration technique called the Full maximum a posteriori method (FMP) [3]. In FMP, we identify a statistical model error for every possible model parameter value. As for many other Bayesian methods, one limitation of FMP is the large number of model solutions that need to be computed and the associated numerical cost. Replacing the numerical code with a surrogate model is one possible and well-known solution.

We propose an adaptive approach for constructing a surrogate model of the computer code that minimizes the approximation error on regions of the parameter domain with high FMP posterior value. Specifically, we build a first surrogate using simulations at a few parameter values and solve the FMP calibration problem. The initial design of experiments is enriched sequentially by drawing new parameter values from the FMP

posterior distribution before updating the surrogate model. As a surrogate, we use the Gaussian process model.

This approach is tested on analytical functions. It is also applied to a complex two-phase flow model implemented in Neptune_CFD solver [2]. More specifically, we calibrate the parameters of the coalescence and breakup processes in the IATE model in the bubbly flow regime, using experimental data from Sun experiment [4]. Results show that the proposed method improves computational efficiency by reducing the number of actual model evaluations required for surrogate construction compared to conventional space filling designs. Additionally, the surrogate we obtain is a reliable approximation of the computer model.

Short biography (PhD student)

Currently, I am a Ph.D. student in the Platon team (CMAP/ INRIA) at Ecole Polytechnique, in partnership with CEA Paris-Saclay. I graduated from ENSEIRB-MATMECA Engineering School with a major in high-performance computing for mechanics. I am working on Bayesian approaches for calibration problems with applications to two-phase flow models.

References

[1] M. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 425–464, 2001.

[2] E.V. Kuidjo Kuidjo, M.G. Rodio, R. Abgrall, and P. Sagaut. Comparison of bubbles interaction mechanisms of two-group Interfacial Area Transport Equation model. *International Journal of Multiphase Flow*, 163:104399, June 2023.

[3] Le-Maître O. Rodio MG. Leoni, N. and PM. Congedo. Bayesian calibration with adaptive model discrepancy. *International Journal for Uncertainty Quantification*, 14:19–41, 2024.

[4] Haomin Sun, Tomoaki Kunugi, Xiuzhong Shen, Dazhuan Wu, and Hideo Nakamura. Upward air–water bubbly flow characteristics in a vertical square duct. *Journal of Nuclear Science and Technology*, 51(3):267–281, March 2014.

Bayesian calibration of a model for predicting the energy consumption of high-speed trains

R. Jorge Do Marco[†,1,2], G. Perrin[§,1], C. Funfschilling[2], C. Soize[§,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Dec. 2022 – Dec. 2025

[1] Université Gustave Eiffel, COSYS, 5 Bd Descartes, 77454 Marne-La-Vallée, France
guillaume.perrin@univ-eiffel.fr

[2] SNCF, Direction Technologies, Innovation et Projets Groupe, 1-3 Av. François Mitterrand, 93574 Saint-Denis, France
{romain.jorge-domarco, christine.funfschilling}@sncf.fr

[3] Université Gustave Eiffel, MSME UMR 8208, 5 Bd Descartes, 77454 Marne-La-Vallée, France
christian.soize@univ-eiffel.fr

Abstract

Reducing ecological impact is a major challenge for today's industry, particularly the rail industry, which is one of the most energy-intensive industries. Indeed, this industry faces two paradoxical needs: on the first hand, it must decrease its energy consumption, meeting both an environmental goal and a financial objective, and on the other hand, it must not only maintain but increase the circulation of train, thus allowing a larger part of the population to use the most ecological means of land transport. Reducing consumption requires a prediction model. The formalization of this model is complex, particularly when driver control is taken into account.

The complexity of the model must be chosen carefully. The SNCF has a very sophisticated model with a large number of parameters that need to be evaluated. In this work, we consider a train dynamics model simplified to a non-linear differential longitudinal dynamics equation coupled to a power balance [1]. In this model it is possible to distinguish two types of inputs: the model parameters which will be calibrated and the environment variables (wind, driver's control, etc.) which will change from one journey to another.

In order to calibrate the model, we rely on a priori knowledge from SNCF experts who provide information on the values of the model parameters. We also have a set of measurements (time, speed, power consumption, control) taken on the same train, for different journeys on different tracks.

The used methodology is Bayesian calibration [2], which makes it possible to use the two types of information available to us while injecting model errors to take account of our imperfect knowledge of the system. These errors are parameterized by hyper-parameters that should also be calibrated.

It is then necessary to formalize the experts' knowledge mathematically in order to create our prior distributions, and then to write a likelihood function that will allow us to take into account both the measurements and the model errors.

Another difficulty is the lack of knowledge about driver control as the control isn't measured. The control linked to each measurement must be determined and appears in our problem as functional hyper-parameters.

The output of this work is a railway application of a known methodology that raises a number of issues. The aim is to use this calibrated model as the basis for learning to optimize quantities such as the energy consumed during a journey.

Short biography (PhD student)

Romain Jorge Do Marco is currently an engineering PhD student at Gustave Eiffel University. He is supervised by Prof. Guillaume Perrin, Prof. Christian Soize and Dr. Christine Funfschilling. His thesis focuses on the real-time optimization under uncertainties of control for train to limit energy consumption. The work is supported by the French railway company SNCF.

References

[1] Julien Nespoulous, Christian Soize, Christine Funfschilling, and Guillaume Perrin. Optimisation of train speed to limit energy consumption. *Vehicle System Dynamics*, 60(10):3540–3557, October 2022.

[2] Christian Soize. *Uncertainty quantification: An accelerated course with advanced applications in computational engineering.* Springer International Publishing, Cham, Switzerland, 2018.

Robustness analysis for UQ using an information-geometric approach: the case of truncated distributions

B. Ketema[†,1,2], F. Costantino[§,1], F. Gamboa[§,1],
R. Sueur [§,2], N. Bousquet[§,2], B. Iooss [§,1,2]

[†] PhD student (presenting author). [§] PhD supervisors

PhD expected duration: Feb. 2023 – Feb. 2026

[1] Institut de mathématiques de Toulouse, Paul Sabatier University,
{baalu-belay.ketema,francesco.costantino,fabrice.gamboa}@univ-toulouse.fr

[2] EDF R&D,
{baalu-belay.ketema,roman.sueur,nicolas.bousquet,bertrand.iooss}@edf.fr

Abstract

In many industrial applications, a computer code denoted G is used to simulate the physical behavior of a system or component. The output $Y = G(X_1, \ldots, X_d)$, sometimes referred to as the system response quantity (SRQ), is a performance indicator or a parameter of the process submitted to operational or regulatory requirements. In a probabilistic uncertainty quantification (UQ) context, the probability distribution of Y represents the uncertainty induced by the uncertain input variables X_k of the problem, which can be affected by various sources. These input uncertainties are usually modeled by parametric distributions among a parametric family $\{P_\theta : \theta \in \Theta\}$.

The perturbed law robustness analysis framework (PL-RA) [4, 3] studies the impact of a perturbation of an input distribution on an output quantity of interest (QoI), which can typically be the α-quantile or super-quantile of Y for a given order α [3]. The Fisher Information defines a metric in the parametric space where θ lies, giving it a Riemannian manifold structure. The derived Fisher-Rao geodesic distance can be used to measure the perturbation applied to input distributions [2]. In most practical situations, the uncertain inputs are bounded. These bounds modify the Fisher geometry of the parametric space compared to the unbounded case.

For instance, univariate Gaussian distributions have the following density function

$$p_{(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

and their Fisher information matrix is given by

$$I_{(\mu,\sigma)} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix},$$

in the (μ, σ) parametrization. This defines a hyperbolic geometry similar to the Poincaré half-plane [1]. If we now consider the truncated family of univariate Gaussian distributions on some interval $[a, b]$, their density function is defined as

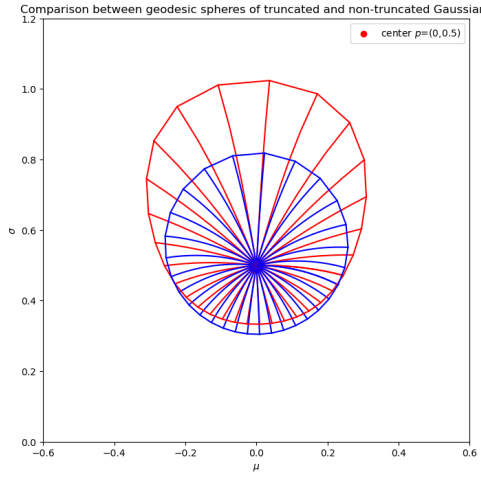


Figure 1: *Comparison of ("small") spheres for the Fisher-Rao distance of the truncated (on* $[-1, 1]$*) and non-truncated Gaussian model. We took* $\theta = (0, 0.5)$ *for the center and* $\delta = 0.5$ *for the radius.*

$$q_{(\mu,\sigma)}(x) = \frac{1}{N_{(\mu,\sigma)}} p_{(\mu,\sigma)}(x) \mathbf{1}_{x\in[a,b]},$$

where $N_{(\mu,\sigma)} = \int_a^b p_{(\mu,\sigma)}(x)\, dx$ is the normalizing constant. In this case, the Fisher information matrix will be different as well as the Riemannian geometric structure it defines. We were able to compute the Fisher matrix analytically as well as the corresponding Christoffel symbols. We were then able to numerically approximate geodesics which allowed us to compute spheres for the corresponding Fisher-Rao distance, see Figure (1).

In this poster presentation, we will expose the general PL-RA framework and illustrate the geometry associated to input random variables with truncated supports.

Short biography (PhD student)

I did my bachelor's and master's degree in mathematics at Paul Sabatier University. I then pursued a PhD career in mathematics at EDF R&D in collaboration with the Mathematics Institute of Toulouse (IMT). The main goal in my thesis is to perform robustness analysis of computer codes for uncertainty quantification using information geometry. I am a second year PhD student and I was funded by the Labex CIMI (Toulouse) for the first few months of my PhD and by EDF for the remaining duration.

References

[1] Sueli IR Costa, Sandra A Santos, and Joao E Strapasson. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015.

[2] Clement Gauchy, Jerome Stenger, Roman Sueur, and Bertrand Iooss. An information geometry approach to robustness analysis for the uncertainty quantification of computer codes. *Technometrics*, 64(1):80–91, 2022.

Robust parameter estimation using variational inference and generative neural networks

E. Ngarti[†,1,2], E. Arnaud[§,1], L. Nicoletti[§,2], V. Trappler[§1,2], A. Vidard[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Mar. 2023 – Mar. 2026

[1] Univ. Grenoble Alpes – Inria – CNRS – LJK, 38000 Grenoble, France
{exauce.luweh-adjim-ngarti,arthur.vidard,elise.arnaud}@inria.fr

[2] AI4Sim R&D team – Eviden (ex Bull/Atos)
leo.nicoletti@eviden.com

Abstract

Estimating key parameters in numerical models is a crucial aspect of numerical simulation, particularly when some parameters are not directly observable. Traditional estimation methods infer these parameters indirectly from their effects on observable variables, thus introducing uncertainties.

Let us consider x as the parameters of interest, y as the observations, and $\mathcal{G}$ as the numerical model. We focus on the posterior distribution $p(x|y)$ of the estimated parameters, which allows for various point estimates such as the mode or the mean, while quantifying the associated uncertainties.

We address the approximation of $p(x|y)$ using variational inference [1], transforming the inference problem into an optimization problem. This involves approximating the posterior distribution with a known family of distributions $\mathcal{Q} = \{q_\phi, \phi \in \Phi\}$, utilizing Kullback-Leibler divergence. It is shown that

$$\min_\phi \mathrm{D}_{\mathrm{KL}}(q_\phi \| p(\cdot|y)) \equiv \min_\phi \mathcal{L}(y;\phi),$$

where

$$\mathcal{L}(y;\phi) := \mathbb{E}_{q_\phi(x)}\left[\log q_\phi(x) - \log p(x,y)\right].$$

Typically, $\mathcal{Q}$ represents a common distribution set such as a Gaussian, and the optimization problem involves selecting optimal Gaussian parameters ϕ^* that minimize $\mathcal{L}(y;\phi)$.

In addition to the parameters to be estimated, there are often uncertain and uncontrollable nuisance parameters in the numerical model. Although variational inference is effective in ideal scenarios, its application in the context of nuisance parameters is less explored. Not accounting for the stochastic nature of these nuisance parameters leads to suboptimal estimation of x due to compensation of errors.

As in [3], these nuisance parameters can be modeled as a random variable $\mathcal{U}$, and the numerical model is redefined as a random variable $\mathcal{G}(x,\mathcal{U})$. Based on this approach, we propose new

strategies for robust variational inference by considering $\mathcal{U}$ in our probabilistic framework. Additionally we integrate generative neural networks such as normalizing flows [2] to enhance the expressiveness of the variational distribution ensemble $\mathcal{Q}$. We applied these methods to the shallow water model, aiming at estimating the friction coefficient, a key parameter in coastal regions that defines the roughness of the seabed. The nuisance parameters represent the boundary forcing due to tidal frequencies and amplitudes.

Short biography (PhD student)

I hold a MSc in Statistics and Stochastic Modeling from Université de Bordeaux and am now a first-year PhD student with Inria's AIRSEA team at Université Grenoble Alpes, collaborating with Evidem's (ex Bull/Atos) AI4Sim R&D team. My thesis focuses on developing robust parameter estimation methods for ocean and atmosphere sciences, funded jointly by Evidem and the French Ministry of Higher Education through a CIFRE PhD program.

References

[1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[2] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[3] Victor Trappler. *Contrôle de paramètre en présence d'incertitudes*. PhD thesis, Université Grenoble Alpes, 2021.

Multifidelity approaches for solving inverse problems relying on computer codes with functional outputs - Application to building envelope thermal performance.

Hadi Nasser[1, †], G. Perrin[1, §], J. Waeytens[1, §], R. Chakir[1, §], S. Demeyer[2, §]

[†] PhD student (presenting author) § PhD supervisor

PhD expected duration: Oct. 2022 – Oct. 2025

[1] Université Gustave Eiffel, IMSE, 5 Bd Descartes - 77447 Marne-la-Vallée, France
Hadi.Nasser@univ-eiffel.fr, Guillaume.Perrin@univ-eiffel.fr,
Julien.Waeytens@univ-eiffel.fr, Rachida.Chakir@univ-eiffel.fr

[2] LNE, 1 r Gaston Boissier, 75015 Paris, France
Severine.Demeyer@lne.fr

Abstract

In the context of energy and environmental renovation, significant advances are expected and necessary in the building sector. For existing buildings, reducing energy consumption requires a better assessment of the energy performance of buildings and its improvement through rehabilitation actions. Special attention must be given to the evaluation and in-situ control of the thermal performance of buildings before and after a rehabilitation action to prevent any defects and thus achieve the expected building performance. Thermal insulation of buildings is a key factor in ensuring the thermal comfort of occupants and achieving high energy savings. Therefore, the control of the thermal resistance of walls (an indicator of the insulation level) through in-situ measurement on existing and new buildings or during renovations is a growing demand. Although many techniques in the literature focus on the building scale [1-2], we have chosen to focus on the scale of walls to propose a solution that is easily deployable, minimally intrusive, and less sensitive to external weather conditions.

The work to be presented focuses on identifying the thermal resistance of highly insulated walls. To accomplish this, a Bayesian sequential multi-fidelity statistical approach will be developed to automatically select the most appropriate version of either the physical model or surrogate model at each iteration of the inversion algorithm. This model represents the best compromise between computational cost and the expected reduction in uncertainty, accurately identifying the thermal resistance of a given wall. To achieve this goal, an original problem formulation will be used to minimize the influence of uncontrolled variables (such as heat exchange coefficients between the wall and the exterior, the initial state of the wall, etc.), and a Bayesian formulation will be applied to incorporate model and measurement uncertainties in the thermal resistance estimation process. This protocol is tested on a 4-layer wall with highly interior insulation, utilizing temperature and flux simulations based on 0D, 1D, and 2D axisymmetric physical wall models.

It can be emphasized that the influence of model error in the inversion process has been studied in previous works [3,4]. Several directions can be explored to minimize the impact of this model error on the results of the inversion at a given total computational cost. For instance, Bayesian sequential strategies inspired by those presented in [5] can be implemented to automatically select the level of fidelity and also the position where to evaluate the code to optimize a criterion for reducing uncertainty on the quantities of interest.

Taking these uncertainties into account, the objective is to avoid biased identification due to overfitting. Furthermore, the use of Bayesian methods will result in a robust confidence interval on the identified thermal resistance.

Short biography (PhD student)

In 2017, I started a bachelor in mathematics and computer science at the University of Montpellier. Then, I realized a master of applied mathematics, statistics at the university of Clermont Auvergne. During the master 1, I did a 5 months research internship at the Magmas et Volcans laboratory. Then, I did a one-year apprenticeship during the master 2, at the Conseil Départemental du Puy de Dôme, in Clermont-Ferrand. Since 2022, I strated a PhD, funded by Agence Nationale de la Recherche, on the subject of identification of the thermal resistance of highly insulated wall, at Univ Gustave Eiffel.

References

[1] S. Roels, Reliable Building Energy Performance Characterisation Based on Full Scale Dynamic Measurements - IEA Annex 58 Seminar – 847, Real building energy perf. assessment, 2014.

[2] F. Alzetto et al., QUB: A fast dynamic method for in-situ measurement of the whole building heat loss - Energy and Buildings, 174, 124-133, 2018.

[3] T-T. Ha et al., Benchmark of identification methods for the estimation of building wall thermal resistance using active method: Numerical study for IWI and single-wall structures - 224, 110-130, Energy and Buildings, 2020b.

[4] J. Waeytens et al., Inverse Computational Fluid Dynamics: Influence of Discretization and Model Errors on Flows in Water Network Including Junctions - J. Fluids Eng., 137, 9, 2015.

[5] G. Perrin, Adaptive calibration of a computer code with time-series output - Reliability Engineering & System Safety, 196, 2020.

Sensitivity analysis for turbulent flows

N. Nouaime[†,1], B. Després[§,2], M-A. Puscas[§,1], C. Fiorini[§,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2021 – Oct. 2024

[1] CEA, DES, ISAS, Modeling of Structure and Systems Department, Center of Saclay, 91191, Gif-sur-Yvette Cedex, France.
nathalie.nouaime@cea.fr

maria-adela.puscas@cea.fr

[2] Sorbonne Université, CNRS, Université de Paris, Laboratoire Jacques-Louis Lions (LJLL), F-75005, Paris, France
bruno.despres@sorbonne-universite.fr

[3] M2N, Conservatoire National des Arts et Métiers, Paris, France.
camilla.fiorini@lecnam.net

Abstract

Sensitivity analysis (SA) studies how changes in the input of a model affect the output. It answers what-if questions, i.e., what happens to the solution of the model if the input parameters change [5]. This task can be performed in many different ways, depending on the nature of the model considered. Our work focuses on sensitivity analysis for the RANS equations more precisely the $k-\epsilon$ model, using the Polynomial Chaos Method (PCM) [3]. PCM is a probabilistic method consisting in the projection of the model output on a basis of orthogonal stochastic polynomials in the random inputs. The stochastic projection provides a representation of the model output variability with regard to the inputs.

PCM has been used for Uncertainty Quantification in many domains. It offers significant advantages in various engineering applications with turbulent flows, despite being a highly challenging problem. Due to the difficulty of applying the PCM to the $k-\epsilon$ [7], we first used it for the Navier-Stokes equations, which is easier to perform.

First, the first-order sensitivity of the Navier-Stokes equations using the PCM is computed. A Finite Element-Volume (FEV) [6, 4] numerical scheme for the Navier-Stokes equations is proposed. This discretisation is integrated into the open-source industrial code TrioCFD [2], developed by the CEA. The FEV method is a widely employed discretisation technique for partial differential equations. It ensures local mass conservation, which is essential for some physical issues.

Second, the first-order sensitivity Navier-Stokes equations are discretised according to the FEV. However, the most significant and original point is the discretisation of the nonlinear term. Then, a stability estimate for continuous and discrete Navier-Stokes equations is established. The PCM provides an estimate of the mean and the variance of the solution of the Navier–Stokes equations when there are uncertain parameters, and this estimated variance is used to compute confidence intervals.

Some numerical tests are presented to evaluate the polynomial chaos method and to compare it to the Monte Carlo and Taylor expansion methods. Finally, we show that the standard PCM

is not directly applicable to the $k - \epsilon$ equations but some novel works are in progress to extend this method to the turbulent flows and the first results are promising.

Short biography (PhD student)

I'm a third-year Ph.D. student; my thesis is entitled "Sensitivity analysis for thermohydrodynamics: uncertainty analysis and estimation of parameters", and the CEA [1] finances it. I'm working on the SA for the incompressible Navier-Stokes equations, discretizing these equations using FEV method, and developing this methods in TrioCFD. Previously, I did a bachelor and a M1 in mathematics at the Lebanese University, and a M2 in modeling and numerical analysis at the University of Montepellier.

References

[1] INSTN. Webpage. https://instn.cea.fr/these/.

[2] P. E. Angeli, M. A. Puscas, G. Fauchet, and A. Cartalade. FVCA8 benchmark for the Stokes and Navier-Stokes equations with the TrioCFD code – benchmark session. In *Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects.*, pages 181–202, 2017.

[3] P. Bonnaire, P. Pettersson, and C.F. Silva. Intrusive generalized polynomial chaos with asynchronous time integration for the solution of the unsteady navier–stokes equations. *Computers and Fluids*, 223:104952, 2021.

[4] Ph. Emonot. *Méthode de volumes éléments finis : applications aux équations de Navier-Stokes et résultats de convergence.* PhD thesis, Université Claude Bernard – Lyon I, 1992.

[5] Camilla Fiorini, Bruno Després, and Maria Adela Puscas. Sensitivity equation method for the navier-stokes equations applied to uncertainty propagation. *International Journal for Numerical Methods in Fluids*, 93(1):71–92, 2021.

[6] S. Heib. *Nouvelles discrétisations non structurées pour des écoulements de fluides à incompressibilité renforcée.* PhD thesis, Université Paris VI, 2003.

[7] Florin Ilinca. *Méthodes d'éléments finis adaptatives pour les écoulements turbulents.* École polytechnique de Montréal, 1996.

Numerical study of the calibration of predictive distributions from Gaussian process interpolation

Aurélien Pion[†,1,2], José Alves[§,2], Emmanuel Vazquez[§,1],

[†] PhD Student. [§] PhD supervisor

PhD expected duration: June 2023 – June 2026

[1] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, Gif-sur-Yvette, 91190

[2] Transvalor S.A.

Abstract

When a function $f : \mathbb{R}^d \to \mathbb{R}$ is observed at points $x_1, \ldots, x_n$, Gaussian Process (GP) interpolation is a standard method to infer the value of f at a new point x. However, what degree of trust can be placed in a prediction from a GP model? The natural way to deal with this question is to study the prediction intervals built from the posterior distribution of the GP model at x. These prediction intervals may be too optimistic or too pessimistic compared to the actual observations (in the precise sense described below). These considerations are crucial in the industry since prediction intervals that are too optimistic can be misleading and have disastrous effects. In contrast, overly pessimistic prediction intervals leads the user to increase the number of observations, resulting in an increasing approximation cost.

In the following, we consider a zero-mean GP model Z with covariance function $\sigma^2\kappa$, where $\sigma > 0$ and κ is a isotropic Matérn covariance kernel with regularity $\nu = p + 1/2$, $p = 1, 2, \ldots$ Given observations $Z(x_1), \ldots, Z(x_n)$, the posterior distribution of Z is again a Gaussian process with posterior mean denoted by $\mu_n(x)$ and posterior covariance function denoted by $\sigma^2\kappa_n$ (see, e.g., [7]).

[4] provides asymptotic results on the coverage of prediction intervals when the number of observations increases in a fixed domain. More precisely, the authors consider prediction intervals written as

$$I_\alpha(x) = \left[\mu_n(x) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sigma p_n(x), \mu_n(x) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sigma p_n(x)\right], \tag{1}$$

where $\alpha \in [0, 1]$ is a probability level, Φ is the cdf of the normal distribution and $p_n(x) = \sqrt{\kappa_n(x, x)}$. They show that the selection of σ by maximum likelihood brings protection against over-optimistic predictions when the GP model is misspecified.

In this work, we conduct a numerical study to check to what extent the results of [4] are verified in a non-asymptotic framework. Moreover, instead of considering the coverage of prediction intervals $I_\alpha(x)$, the metric we use is the variance of the *Probabilistic Integral Transform* (PIT).

The PIT is a tool widely used in the meteorology community to study the *calibration* of predictive distributions from a given prediction method (see[1], [2] or [3]). In our numerical experiments, the PIT is computed on a test set. A random Latin Hypercube Sampling (LHS) method is employed to sample observation points of the training set $\{x_i^{\text{train}}\}$, and a separate LHS is used for the points of the test set $\{x_i^{\text{test}}\}$. We assess the properties of predictive distributions obtained from a

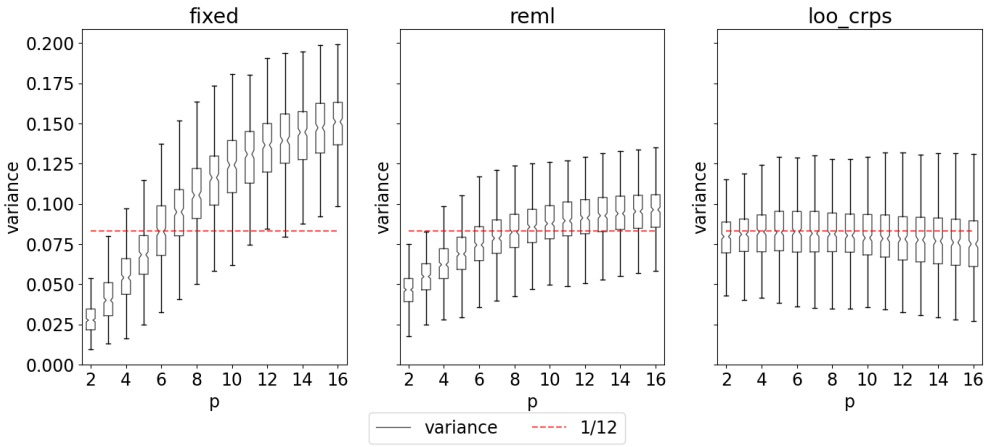


Figure 1: Boxplots of the variance of the probability integral transform (PIT) for a prediction model with varying regularity ($p = 2, 3, \ldots$). The parameters for the generated sample paths are $\sigma_0^2 = 0.3$, $1/\rho_0 = 1.39$, $p_0 = 8$, and $d = 3$. The red lines indicate a variance of $\frac{1}{12}$, which corresponds to *neutrally dispersed* predictions, as explained in [2]. On the left-hand side, the parameter σ in model Z is fixed at $\sigma = \sigma_0$. In the center, σ is selected using maximum likelihood estimation. On the right-hand side, σ is chosen based on leave-one-out cross-validation using a CRPS score, as detailed in [6].

model $Z \sim \mathrm{GP}(0, \sigma^2 \kappa_{\rho,p})$ using sample paths of a GP $Z_0 \sim \mathrm{GP}(0, \sigma_0^2 \kappa_{\rho_0,p_0})$ on the domain $[0,1]^d$. More precisely, for each sample path f_i and test point x_j^{test}, we compute the posterior cumulative distribution functions (CDF) $F_{i,x_j^{\text{test}}}$ of Z conditioned on the training set $\{x_j^{\text{train}}, z_{i,j}^{\text{train}}\}$, where $z_{i,j}^{\text{train}} = f_i(x_j^{\text{train}})$. The PIT is defined as the sample $\left\{F_{i,x_j^{\text{test}}}\left(z_{i,j}^{\text{test}}\right)\right\}_{j \in [\![1, n^{\text{test}}]\!]}$, where $z_{i,j}^{\text{test}} = f_i(x_j^{\text{test}})$. If the predictions are well-calibrated, meaning neither overly optimistic nor excessively pessimistic, the empirical distribution of the PIT is close to the uniform distribution. In such instances, the PIT's variance is approximately equal to $\frac{1}{12}$. The predictions are pessimistic when the variance is below $\frac{1}{12}$, and the predictions are optimistic when the variance is above $\frac{1}{12}$. The boxplots in Figure 1 depict the variances of the PIT for the sample paths f_i, each boxplots is associated with varying regularity values p of the underlying model Z. In this particular experiment, we study the calibration properties for different methods of parameter selection. In this work, our numerical experiments investigated several methods of parameter selection, the influence of the size of the design, and generalization properties to see if we can infer calibration properties from leave-one-out procedures.

This study provides a better understanding of the behavior of parameter selection methods for Gaussian processes in terms of calibration of prediction intervals. Note that Amandine Marrel and co-authors [5] have recently proposed a method for obtaining better calibration. This topic is outside the scope of this work, but is a very interesting avenue for future research.

Short biography (PhD student)

I started my PhD thesis in June 2023. Before that, I graduated from the engineering school l'Ecole des Ponts in 2023, where I specialized in statistics and machine learning with the master "Mathématiques, Vision, Apprentissage" (MVA) from l'ENS Paris-Saclay in 2022. I interned in 2022 at the Laboratoire des Signaux et Système under the supervision of Emmanuel Vazquez and in collaboration with Transvalor S.A. to prepare the thesis. Transvalor S.A. is a company specializing in developing simulation software for the industry. In 2023, an industrial research

An automated surrogate modelling approach for the uncertainty quantification of dynamical systems

S. Schär[†,1], S. Marelli[§,1], B. Sudret[§,1],

[†] PhD student [§] PhD supervisor

PhD expected duration: Apr. 2021 – Apr. 2025

[1] Chair of Risk, Safety & Uncertainty Quantification, ETH Zürich
{styfen.schaer,marelli,sudret}@ibk.baug.ethz.ch

Abstract

Modelling dynamical systems plays a pivotal role in numerous disciplines in engineering and applied science. It can help deepen our understanding of system dynamics, predict the future state of a system from past observations and measurements, and reduce the costs of system response analysis. The latter is achieved by building a fast surrogate for the system under investigation. This is crucial in uncertainty quantification tasks, where many evaluations of complex computational models, such as finite element models, are often required.

A wide variety of surrogate models can map a set of scalar input parameters to one or multiple scalar output quantities. For a dynamical system $\mathcal{M}$, however, the system response $y(t) \in \mathbb{R}$ is a time series that is generally driven by a temporally coherent exogenous excitation $\boldsymbol{x}(t) \in \mathbb{R}^M$, as follows:

$$y(t) = \mathcal{M}(\boldsymbol{x}(\mathcal{T} \leq t), \boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{\beta}$ represents a set of initial conditions and $\boldsymbol{x}(\mathcal{T} \leq t)$ denotes that the output at time t depends on the input up to and including time t.

To surrogate such systems, a powerful option is given by nonlinear autoregressive with exogenous inputs (NARX) [1]. A NARX model $\tilde{\mathcal{M}}$ predicts the near-future time evolution of the model response based on current and past inputs, as well as past outputs:

$$y(t) = \tilde{\mathcal{M}}(\boldsymbol{x}(\mathcal{T} \leq t), y(\mathcal{T} < t)). \tag{2}$$

While this allows NARX models to take advantage of temporal dependencies in the data, it also makes them highly susceptible to the curse of dimensionality. This is particularly the case when dealing with highly non-linear system responses, where many past time steps are needed to describe the system dynamics, or in cases where the system response depends on many input features.

To address the challenge of surrogating these complex dynamical systems with high input dimensionality, we recently developed a novel approach called manifold NARX (mNARX) [2]. mNARX involves the assembly of an exogenous input manifold, specifically designed to simplify the construction of autoregressive surrogate models. The manifold consists of incrementally constructed auxiliary quantities, which integrate the physical principles of the system being modelled, as well as expert and domain-specific knowledge. This incremental construction breaks down the entire modelling problem into smaller, more manageable sub-problems. Each sub-problem is addressed using a conventional NARX model, which is typically of relatively low complexity, allowing mNARX to scale favorably with increasing problem complexity. Consequently, mNARX is particularly suitable for systems characterized by complex dynamics, such as coupled systems or systems manipulated by control units. Additionally, its compatibility with standard dimensionality reduction methods further enhances its applicability in modelling dynamical systems with high-dimensional exogenous inputs.

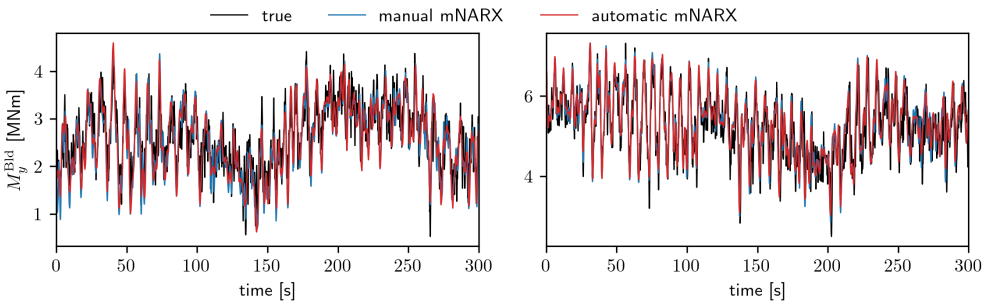


Figure 1: Visual comparison of two exemplary traces of the flapwise blade root bending moment (M_y^{Bld}) from the *manually* and *automatically* build mNARX surrogates.

Using prior system knowledge to construct the exogenous input manifold renders mNARX extremely efficient for various engineering problems. However, identifying which auxiliary quantities to construct and determining their order of construction can be challenging in the general case. This complexity may hinder the adoption of mNARX by a less specialized audience, or in situations with limited information about the system under investigation. Moreover, mNARX shares a characteristic with traditional NARX models: sensitivity to its structure. Selecting a suitable model structure for multiple NARX models can be a challenging process. To overcome these challenges, we enhanced the original mNARX algorithm with an additional step that can automatically select relevant auxiliary quantities from a pool of candidates, as well as determine their optimal construction order, using a recursive correlation-based algorithm. Since fitting a NARX model can be formulated as an ordinary regression problem, we also use sparse regression techniques to facilitate the NARX model structure selection.

We evaluate the enhanced mNARX algorithm by comparing it to the original version on an onshore wind turbine simulator case study. The objective is to build a surrogate for the simulator that produces various response quantities $f_i(t) \in \mathbb{R}$, such as blade root bending moment or power production, based solely on an input wind field $\boldsymbol{v}(t) \in \mathbb{R}^M$ with $M = \mathcal{O}(10^3)$:

$$f_i(t) = \tilde{\mathcal{M}}(\boldsymbol{v}(\mathcal{T} \leq t)), \tag{3}$$

The response of the turbine is also influenced by multiple other factors, such as the turbine state (e.g., rotor orientation), and its control system, which regulates certain degrees of freedom such as the blade pitch angle. This results in a highly non-linear input-output mapping, making it a challenging problem that is ideal for assessing the performance of the mNARX algorithm. A visual comparison of the flapwise blade root bending moment produced from both manually built and automatically constructed mNARX surrogates is shown in Fig. 1. The figure shows that the automatic construction of the surrogate results in a model with comparable performance.

Short biography

Styfen Schär is a Ph.D. student at the Chair of Risk, Safety and Uncertainty Quantification at ETH Zürich. He is developing new surrogate modelling techniques for predicting the long-term responses of complex dynamical systems with high-dimensional exogenous inputs.

References

[1] S. A. Billings. *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons, Inc, Chichester, West Sussex, United Kingdom, 2013.

[2] S. Schär, S. Marelli, and B. Sudret. Emulating the dynamics of complex systems using autoregressive models on manifolds (mNARX). *Mechanical Systems and Signal Processing*, 208:110956, 2024.

Clouds of points optimization in convex polygons with evolutionary algorithms based on Wasserstein barycenters

B. SOW[†,1], R. LE RICHE[§,2], J. PELAMATTI[§,3], M. KELLER[§,3], S. ZANNANE[§,3]

[†] PhD student (presenting author), babacar.sow@emse.fr
[§] PhD supervisors

PhD expected duration: Nov. 2021 – Oct. 2024

[1] LIMOS, Ecole Nationale Superieure des Mines de SAINT-ETIENNE
[2] CNRS/LIMOS [3] EDF R&D

Abstract

We consider the optimization of functions where the variables are clouds of points (equivalently bags of vectors). The clouds can have different sizes and the objective functions are invariant under arbitrary permutation of the points within the cloud. Furthermore, no information related to the convexity and/or smoothness of the functions is available. Such functions are of practical interest as they appear, for example, when studying networks of sensors or actuators. The design of a wind farm where the variables are a bag of turbine positions belongs to this family of problems.

The above characteristics make it hard to use off-the-shelf algorithms such as gradient-based methods. We introduce a generic approach for such black-box optimization problems where the cloud is modeled as a discrete uniform measure supported by the cloud points. This allows to define stochastic, evolutionary, optimization algorithms whose transition operators (crossover and mutation) use Wasserstein barycenters (see [1] for a review). The crossover operator interpolates between two clouds of points, by calculating their average in the Wasserstein sense. We give an illustration of such an interpolation on clouds in Figure 1. We prove that the Wasserstein barycenter has a contracting effect, which can lead to a premature convergence to degenerated solutions. Specific mutations, once again based on Wasserstein barycenters, are designed to counteract this effect.

We first investigate the performance of variants of our Wasserstein-based optimizers by comparing them to classical evolutionary algorithms on a family of test functions including wind farm proxies. The point domains are convex polygons such as squares and trapezes. Second, the points optimization domain is generalized to involve exclusion zones, defined as subdomains which must not include any point of the optimal solution. Exclusion zones constitute a specific type of constraint which is handled by penalization.

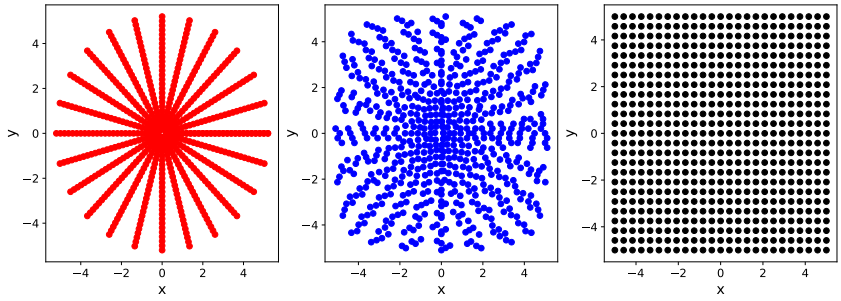


Figure 1: Left (in red) and right (in black) are two clouds of points. Their equally weighted Wasserstein barycenter is in the center (in blue).

Short biography (PhD student)

Babacar Sow has received his engineering degree from Telecom Paris and Masters degree from Institut Polytechnique de Paris in Data Science And Applied Statistics. He currently pursues a PhD with Mines Saint-Etienne in collaboration with EDF. This work is funded by the ANR SAMOURAI project ANR-20-CE46-0013.

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Equivariances of random field paths with applications in Gaussian process modelling

T. Steinert[†,1], D. Ginsbourger[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2023 – Sep. 2027

[1] Institute of Mathematical Statistics and Actuarial Science, Department of Mathematics and Statistics, University of Berne
tim.steinert@unibe.ch

Abstract

We study the incorporation of equivariances in vector-valued random field models relying on equivariant matrix-valued kernel functions as introduced by [4]. In [2], authors show that for broad classes of invariances in scalar valued settings, invariance of centered second-order random fields can be characterized by their associated covariance kernel being (argumentwise) invariant. We extend these results to vector-valued settings for a class of equvariances following the work in [4].

Proposition 1. *Let $Z = (Z_x)_{x\in \mathrm{R}^d}$, be a square-integrable centred random field taking values in R^p, with kernel function $K : \mathrm{R}^d \times \mathrm{R}^d \to \mathrm{R}^{p\times p}$. Furthermore let $\mathcal{G}$ be a compact, linear, unimodular group with representation $\rho_g \in \mathrm{R}^{p\times p}$ for any $g \in \mathcal{G}$. Then, the random field Z is $\mathcal{G}-$equivariant up to a modification if and only if the corresponding kernel matrix function K is $\mathcal{G}-$equivariant, i.e.*

$$\forall x \in D,\ g \in \mathcal{G},\ \ \mathbb{P}(Z_{g.x} = \rho_g Z_x) = 1 \iff K(g.x, h.x') = \rho_g K(x, x')\rho_h^T \ \ \forall x, x' \in D,\ g, h \in \mathcal{G}.$$

Following [4], under the assumptions of the proposition and assuming in case of a continuous group $\mathcal{G}$, $g \mapsto \rho_g$ to be continuous, such an equivariant matrix-valued kernel function can be constructed by Haar integration of a base kernel matrix function K with respect to $\mathcal{G}$, i.e.

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = \int_{\mathcal{G}^2} \rho_g^T K(g.x, h.x')\rho_h \ dg \ dh.$$

It is possible to reduce the double integration to a single integration by using an invariant base kernel matrix function, i.e. $K(g.x, g.x') = K(x, x')$.

We compare the performance of the Gaussian process with baseline squared-exponential velocity kernel-matrix function $K(x, x') = diag(\sigma_i^2 \exp(-||x - x'||^2/2l_i^2),\ i \in \{1, 2\}),\ x, x' \in \mathrm{R}^2$, with its rotation-equivariant version on ocean velocity vector fields as seen in [1]. We consider two two-dimensional ocean velocity vector fields over two-dimensional locations (here the domain is restricted to a square), the first being computed as $F_1(x) = (-x_2, x_1)$ and the second as $F_2(x) = x/(0.4 + ||x||^4)$. We observe the ocean velocities at 8 (resp. 10) locations with noisy measurements $F(x) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_{obs}^2 I_2)$ and $\sigma_{obs}^2 = 0.2$ (resp. $\sigma_{obs}^2 = 0.1$). We estimate the parameters l_i, σ_i^2, σ_{obs}^2 by maximum likelihood and predict the ocean velocities on a 17×17 (resp. 20×20) equidistant grid on $[-1, 1]^2$ (resp. $[-2, 2]^2$). The results are shown in Figure 1.

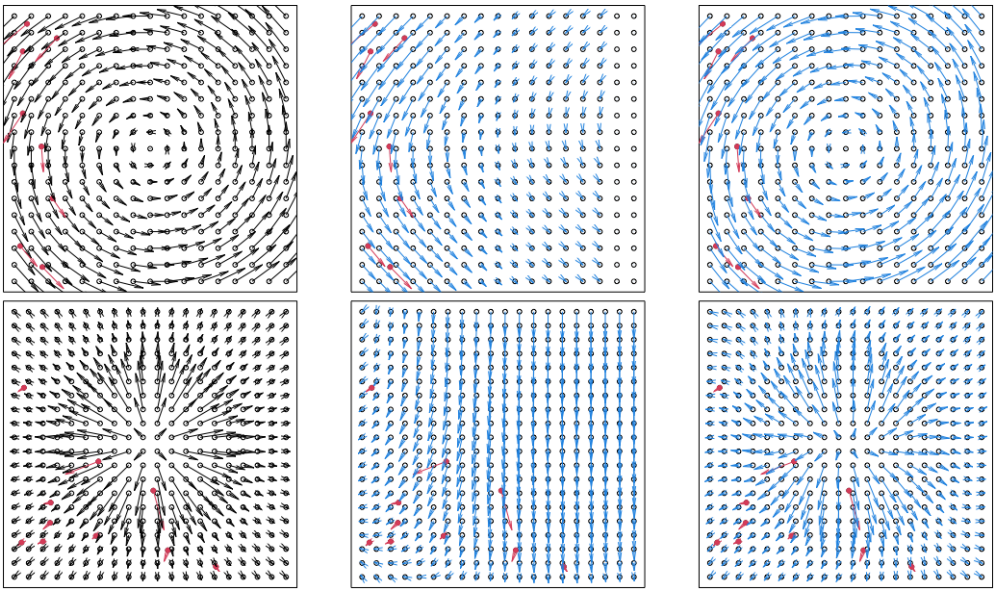


Figure 1: Ground truth (left), predicted velocities using standard SE-velocity kernel (middle) and predicted velocities using equivariant SE-velocity kernel (right). The top row corresponds to the first simulation, the bottom row to the second and red locations indicate (noisy) observations.

We observe that the posterior mean of the equivariant GP is able to predict the left out data much better than the baseline GP. For F_1, we obtain RMSE values of 0.12 (equivariant kernel) versus 0.65 (baseline kernel) and for F_2, 0.16 (equivariant) versus 0.85 (baseline). The results can be reproduced with the code on Github.

The authors are thankful for Henry Moss (Cambridge) as our preliminary discussions have lead to this work.

References

[1] Renato Berlinghieri, Brian L. Trippe, David R. Burt, Ryan Giordano, Kaushik Srinivasan, Tamay Özgökmen, Junfei Xia, and Tamara Broderick. Gaussian processes at the helm(holtz): A more fluid model for ocean currents, 2023.

[2] David Ginsbourger, Olivier Roustant, and Nicolas Durrande. On degeneracy and invariances of random fields paths with applications in gaussian process modelling. *Journal of Statistical Planning and Inference*, 170:117–128, 2016.

[3] Iain Henderson. *PDE constrained kernel regression methods*. PhD thesis, INSA Toulouse; Université de Toulouse-Toulouse III-UPS, 2023.

[4] Marco Reisert and Hans Burkhardt. Learning equivariant functions with matrix valued kernels. *Journal of Machine Learning Research*, 8(15):385–408, 2007.

Short biography (PhD student)

I am a first year PhD student with a background in Applied Mathematics. My PhD studies about probabilistic kernel methods under the project "Perception in Statistics and Econometrics" are being funded by the program "Mensch in digitaler Transformation" of University of Berne.

Robust design of experiment based on sensitivity indices for seismic fragility curves estimation

A. Van Biesbroeck[†,1,2], C. Gauchy[§,3], C. Feau[§,2], J. Garnier[§,1]

[†] PhD student (presenting author). [§] PhD supervisors

PhD expected duration: Oct. 2022 – Sep. 2025

[1] CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, France
antoine.van-biesbroeck@polytechnique.edu

[2] Université Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, France

[3] Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, France

Abstract

Seismic fragility curves of mechanical structures are key quantities of interest for probabilistic seismic risk assessment studies. They express the probability of failure of the mechanical structure conditional to a scalar value derived from the seismic ground motion, coined intensity measure. Estimating such curves is a daunting task because, in numerous case studies the number of data available is limited. For this reason, a wide range of the methods of the literature relies on the parametric log-normal model.

This work targets studies for which the information on the response of the structure is limited to binary information of failure or not, for a given ground motion signal. We are particularly interested in experimental results obtained on complex systems, such as electrical devices, that cannot be modeled numerically.

For those, the Bayesian approach permits an efficient learning of the parameters which determine the fragility curves. It avoids the generation of unrealistic fragility curves such as unit step functions, which are common with classical methods. Nevertheless, it can still be affected by the occurrence of what we call degenerative scenarii. They are related to the distribution of the experimental results and can lead to improper posterior distributions [5]. The probability of occurrence of these scenarii is affected by different factors such as (i) the number of data, (ii) the failure threshold of the structure and (iii) the intensity measure.

We therefore propose a design of experiment methodology, based on the measure of the posterior's sensitivity to the observed data, and supported by generalized sensitivity indices built with f-divergences [2]. The prior selection is also questioned as it is known that its choice may have a significant influence on the posterior distribution when the dataset is small. Our prior construction is based on the reference prior theory to support the objectivity of our choice [1, 4]. The prior distribution is derived from the asymptotic of the Jeffreys' prior in the same fashion as in [3].

The consistency of the method is evaluated on different case studies, and our results prove its capacity to efficiently solve the problem of degenerative sample generation, enhancing any estimation of the seismic fragility curves.

Short biography (PhD student)

The PhD student has a background that mixes both industrial and academic experiences. This PhD allows a work within those two areas, with the idea of providing theoretical tools from the Bayesian statistics field to the enhancement of the estimation of the seismic fragility of industrial facilities. It is motivated by the CEA's needs which funds this work and provides simulated or experimental data to evaluate it.

References

[1] James O. Berger, José M. Bernardo, and Dongchu Sun. The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938, 05 2009.

[2] Sébastien Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, 2015.

[3] Mengyang Gu, Xiaojing Wang, and James O. Berger. Robust Gaussian stochastic process emulation. *The Annals of Statistics*, 46(6A):3038–3066, 2018.

[4] Antoine Van Biesbroeck. Generalized mutual information and their reference prior under Csizar f-divergence. 2023. arXiv.2310.10530.

[5] Antoine Van Biesbroeck, Clément Gauchy, Cyril Feau, and Josselin Garnier. Reference prior for Bayesian estimation of seismic fragility curves. 2023. arXiv.2302.06935.

Diffeomorphism-based feature learning using Poincaré inequalities on augmented input space

R. Verdière[†,1], C. Prieur[§,1] O. Zahm[§,1],

[†] PhD student (presenting author). [§] PhD supervisor
PhD expected duration: Sept. 2022 – Aug. 2025

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK
{romain.verdiere,olivier.zahm}@inria.fr ; clementine.prieur@univ-grenoble-alpes.fr

Abstract

Modern computational models for scientific and engineering applications typically involve a large number of input parameters and are expensive-to-evaluate both in time and resources. Replacing the model with an accurate and fast-to-evaluate surrogate (or approximation) offers a viable workaround in many applications. Approximating such high-dimensional functions with classical approximation tools such as polynomials, wavelets or neural networks is, however, a difficult task. This is even aggravated in the small sample regime where one only has access to a little number of model evaluations. One way to address this challenge is to reduce the input dimension beforehand. This consists in approximating the *model* (u) as the composition of two functions: a *feature map* (g) which extracts the relevant features of the input variables, and a *profile function* (f) which regresses the model output on the features.

In this talk we propose a gradient-enhanced algorithm for high-dimensional function approximation. The algorithm proceeds in two steps: firstly, we reduce input dimension by learning a feature map g from gradient evaluations, and secondly, we regress the model output $u(\mathbf{X})$ against pre-learned features $g(\mathbf{X})$. Specifically, we build a nonlinear feature map as the first components of a diffeomorphism defined on input space, a solution originally considered in [4]. Based on this diffeomorphism structure, we pursue the analysis of [1] by applying Poincaré inequalities on either input space or feature space, defined as the range of the diffeomorphism. That way, we obtain two different error bounds ($\mathcal{J}_m^{\mathrm{IS}}(g)$ or $\mathcal{J}_m^{\mathrm{FS}}(g)$) which we minimize in order to train a feature map. Let us note that such a strategy of optimizing an error bound is a well-tried strategy in many machine learning problems, see for instance the use of the evidence-based lower bound for optimizing variational auto-encoders. By considering such diffeomorphism-based feature maps, however, we drastically restrict the approximation class for features. To circumvent this issue, we propose a dimension augmentation strategy to increase the expressiveness of features while preserving theoretical foundation of the method. The basic idea, originally proposed for neural ODEs [2], is to introduce a new arbitrary random variable $\boldsymbol{\Xi}$ taking values in $\Xi \subseteq \mathbb{R}^k$, and to augment input vector as $(\mathbf{X}, \boldsymbol{\Xi})$. Using this dimension augmentation trick, we no longer approximate $u(x)$ by a composition $f \circ g(x)$ but rather by a random variable $f \circ g(x, \boldsymbol{\Xi})$.

In practice, we construct the diffeomorphism, and consequently the feature map, using *coupling flow* based neural networks. This class of invertible networks is known for achieving favorable approximation properties. Numerical comparisons with existing methods from the literature in [1, 4, 3] are presented. Our numerical demonstrations highlight the efficiency of our methodology in achieving accurate approximations across various high-dimensional test cases. Notably, in the small sample regime, employing dimension augmentation outperforms existing state-of-the-art methods.

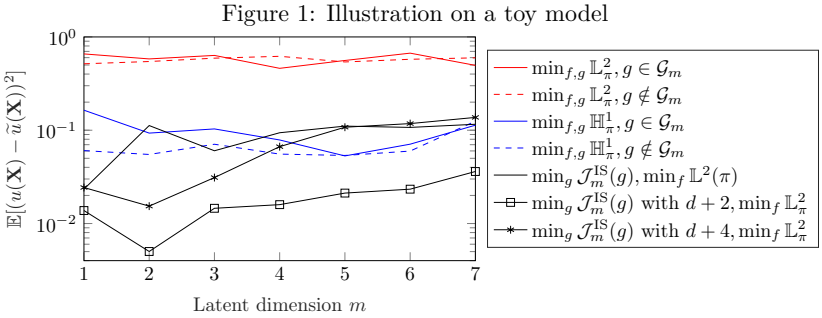
Figure 1 demonstrates numerically the benefit of the strategy we propose on a toy model. For this experiment, we considered $u(\mathbf{X}) = \cos(\|\mathbf{X}\|^2)$ for $\mathbf{X} \sim \mathcal{U}([0,1]^d)$ with $d = 20$ where $\|\cdot\|$ is the

euclidean norm of $\mathbb{R}^d$. The approximation class for g is defined as

$$\mathcal{G}_m = \left\{ \begin{array}{rccl} g: & \mathcal{X} & \to & \mathbb{R}^m \\ & x & \mapsto & (\varphi_1(x), \ldots, \varphi_m(x)) \end{array} \middle| \, \varphi \in \mathcal{D} \right\}$$

with $\mathcal{D}$ a set of diffeomorphisms. More precisely, $g \in \mathcal{G}_m$ is a block affine coupling flow with 4 layers. The different curves on Figure 1 represent $\mathbb{L}^2_\pi$ error as a function of latent (reduced) dimension m, where $\mathbb{L}^2_\pi$ error is defined as $\mathbb{E}[(u(\mathbf{X}) - \widetilde{u}(\mathbf{X}))^2]$, with $\widetilde{u}(x) = f \circ g(x)$ or, when dimension augmentation is used, $\widetilde{u}(x) = \mathbb{E}[f(g(x, \boldsymbol{\Xi}))]$. Black curves correspond to the two-step strategy, minimizing first the bound $\mathcal{J}^{\mathrm{IS}}_m(g)$ obtained applying Poincaré Inequality in input space without (solid) or with dimension augmentation $d + k$, $k \in \{2, 4\}$ (squares, stars), and then learning f as a fully connected neural network with 3 hidden layers of 50 neurons using the sigmoid activation function. Red and blue curves were obtained by training f and g jointly by minimizing the $\mathbb{L}^2_\pi$ error (red) or the $\mathbb{H}^1_\pi$ error (blue), with either $g \notin \mathcal{G}_m$ (dashed) or $g \in \mathcal{G}_m$ (solid). For $g \notin \mathcal{G}_m$, the feature map is a fully connected neural network with 3 hidden layers of 26 neurons for an overall parameter number comparable from one strategy to the other.

Figure 1: Illustration on a toy model



Short biography

I am a former student of ENSTA Paris where I studied mathematics and computer science. At the end of my engineering studies, to deepen my knowledge in general mathematics, I prepared for and passed mathematical aggregation. Finally, after a short experience as a teacher, I moved to Grenoble to work as PhD student with Clémentine Prieur and Olivier Zahm on nonlinear dimension reduction for function approximation. This thesis is funded by the French Research Agency (ANR) and this communication summarizes the preprint available here: `https://hal.science/hal-04364208`.

References

[1] Daniele Bigoni, Youssef Marzouk, Clémentine Prieur, and Olivier Zahm. Nonlinear dimension reduction for surrogate modeling using gradient information. *Information and Inference: A Journal of the IMA*, 11(4):1597–1639, 2022.

[2] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019.

[3] Yuankai Teng, Zhu Wang, Lili Ju, Anthony Gruber, and Guannan Zhang. Level set learning with pseudo-reversible neural networks for nonlinear dimension reduction in function approximation. *arXiv preprint arXiv:2112.01438*, 2021.

[4] Guannan Zhang, Jiaxin Zhang, and Jacob Hinkle. Learning nonlinear level sets for dimensionality reduction in function approximation. *Advances in Neural Information Processing Systems*, 32, 2019.

Sensitivity analysis for time-varying systems

M. Yachouti[†,1], J. Garnier[§,1], G. Perrin[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Jan. 2024 – Jan. 2027

[1] CMAP, Ecole polytechnique, Institut Polytechnique de Paris
{mouad.yachouti, josselin.garnier}@polytechnique.edu

[2] COSYS, Université Gustave Eiffel
guillaume.perrin@univ-eiffel.fr

Abstract

A vast amount of historical data is collected to monitor complex dynamic systems in different fields, such as physics, biology, finance, manufacturing, etc. Often, we are interested in following the good functioning (or, alternatively, the failure) of a system based on a key quantity y. Correspondingly, many variables characterizing its environment and may have influence on its behavior can be monitored. Therefore, there is a critical need for the analysis and the assessment of the risk related to these time varying systems.

Let $t \mapsto y(t)$ and $t \mapsto \boldsymbol{x}(t)$ denote respectively the temporal evolution of the response and the environment variables of some system. The first challenge is to achieve sensitivity analysis [1], ie. to detect which variables actually have an important influence on y and to quantify it, as well as those towards which y has little or no sensitivity. Answering this question is non-trivial due to the functional nature of the variables and the response [2], the potential temporal correlation of the components of $\boldsymbol{x}$ (for example, in the presence of a daily periodicity in the data), or the fact that the impact of their evolution at a time t is observed on y at a time $t + \tau$ for a non-negligible response time τ.

We can then ask to what extent the knowledge acquired on the couple $(\boldsymbol{x}, y)$ at times $t \leq \tau$ allows to estimate the probability of a feared event on y at a time $\tau' > \tau$ (for example, the probability that y exceeds a certain threshold S at τ'). It may also be important, for this estimation, to take account of the possible dependence of this event on latent variables that are not included in $\boldsymbol{x}$. Then, for risk explainability and analysis purposes [3], it may be interesting to characterize the distribution of $\{\boldsymbol{x}(t),\, t \leq \tau\}$ conditioned by the fact that $y(\tau) > S$. Looking at the most possible trajectories resulting from this distribution should perhaps enable the design of representative and plausible critical scenarios.

The aim of the thesis is to propose a methodology to address the aforementioned interrogations. The methods developed during this work will be applied, first, to realistic synthetic data, and then, to real-world application cases.

Short biography (PhD student)

To complete his Master's in applied mathematics from Université Paris Cité, Mouad did an end-of-studies internship on explainable artificial intelligence at Michelin. He then joined the Centre Borelli as a research assistant in Machine Learning, before starting his PhD thesis within the framework of the BNP Paribas stress test chair at CMAP - Ecole polytechnique.

References

[1] Sébastien Da Veiga, Fabrice Gamboa, Bertrand Iooss, and Clémentine Prieur. *Basics and Trends in Sensitivity Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021.

[2] Fabrice Gamboa, Alexandre Janon, Thierry Klein, and Agnès Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8(1):575 – 603, 2014.

[3] Marouane Il Idrissi, Vincent Chabridon, and Bertrand Iooss. Developments and applications of shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling Software*, 143:105115, 2021.

Greedy density estimation using Sum-of-Squares

B. Zanger[†,1,2], O. Zahm[§,1,2], M. Schreiber[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Sep. 2022 – Sep. 2025

[1] INRIA Grenoble Rhône-Alpes, France
{benjamin.zanger, olivier.zahm}@inria.fr

[2] Université Grenoble Alpes, France
{martin.schreiber}@univ-grenoble-alpes.fr

Abstract

Transport-based density approximation methods are receiving growing interest, because of their ability to computationally cheaply sample from the approximated distribution. These methods use a transport map $\mathcal{T}$ to transport a reference probability distribution ρ_{ref} to a probability distribution to be approximated, which we call the target distribution π_f and is proportional to a measure f, $\pi_f \propto f$ (taken w.r.t the Lebesgue measure). Ultimately, the goal of transport based density estimation methods is to find a map $\mathcal{T}$ so that $\mathcal{T}_\sharp\, \rho_{\text{ref}} \approx \pi_f$ for some known ρ_{ref}. Typically, this is done by using variational methods together with a statistical distance D and to minimize

$$\min_{\mathcal{T}_\sharp\, \rho_{\text{ref}} \in \mathcal{M}} \mathrm{D}(\pi_f || \mathcal{T}_\sharp\, \rho_{\text{ref}}) \tag{1}$$

over a set of probability densities $\mathcal{M}$. For instance, Normalizing flow [4] consider $\mathcal{M} = \{(\mathcal{T})_\sharp\, \rho_{\text{ref}} : \mathcal{T} \in \text{Diffeo}\}$ where $\mathcal{T}$ (and sometimes ρ_{ref}) is parameterized and learned by neural networks. Another method is to learn an intermediate density which can be used to analytically build $\mathcal{T}$, e.g. [3] and [5] propose $\mathcal{M} = \{\mathcal{T}_\sharp\, \rho_{\text{ref}} = \widetilde{\pi} \propto p^2 \mu : p \in L^2_\mu\}$, where μ denotes some known measure, and $\mathcal{T}$ can be analytically build by the approximation using the Knothe-Rosenblatt rearrangement.

In our work, we propose to use Sum-of-Squares (SoS) for approximating densities, given by$\mathcal{M} = \left\{\Phi^T A \Phi | A \succcurlyeq 0\right\}$, and use the Knothe-Rosenblatt rearrangement to build $\mathcal{T}$. SoS posses inherent positivity and optimization over convex D remains convex. Since in general the variational problem is difficult to solve, in particular when the density π_f is concentrated on a manifold in high-dimension, we propose to resolve that issue by considering a sequence of bridging densities

$$\pi_{f^{(1)}}, \pi_{f^{(2)}}, \ldots, \pi_{f^{(L)}} = \pi_f, \tag{2}$$

with increasing complexity. Following the ideas of greedy approximations, proposed in [1, 2] as *deep* and in [3] as *self-reinforced*, we build approximations $\pi_{\tilde{f}^{(\ell)}}$ of $\pi_{f^{(\ell)}}$ in the form of

$$\pi_{\tilde{f}^{(\ell)}} = (\mathcal{T}_\ell)_\sharp\, \rho_{\text{ref}}, \quad \text{where } \mathcal{T}_\ell = \mathcal{Q}_1 \circ \ldots \circ \mathcal{Q}_\ell, \tag{3}$$

where each diffeomorphic map $\mathcal{Q}_\ell$ is constructed one after the other such that

$$(\mathcal{Q}_\ell)_\sharp\, \rho_{\text{ref}} = \pi_{\tilde{g}^{(\ell)}} \propto \tilde{g}^{(\ell)}, \tag{4}$$

and $\pi_{\tilde{g}^{(\ell)}} \in \mathcal{M}$ solves the variational problem

$$\min_{\pi_{\tilde{g}^{(\ell)}} \in \mathcal{M}_{\mathrm{SoS}}} \mathrm{D}((\mathcal{T}_{\ell-1})^\sharp \pi_{f^{(\ell)}} || \pi_{\tilde{g}^{(\ell)}}). \tag{5}$$

An example of such an approximation using 3 greedy steps is depicted in Figure 1.

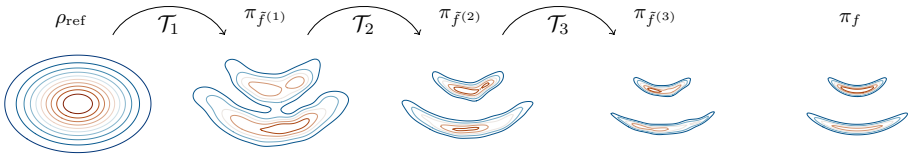


Figure 1: Visualization of the approximation of a double banana distribution π_{tar} using 3 self reinforced layers with a guassian reference distribution ρ_{ref}.

For the statistical distance in Eq. (5), we propose to use the class of α-divergences D_α for D. Instead of working with probability densities π_f, we work with measures (unnormalized densities) f, for which the α-divergences are well defined. Hence, we change the problem of variational density estimation in Eq. (5) to

$$\min_{\tilde{g}^{(\ell)} \in \mathcal{M}} \mathrm{D}_\alpha \left((\mathcal{T}_{\ell-1})^\sharp f^{(\ell)} || \tilde{g}^{(\ell)} \right) \tag{6}$$

where the estimation of $\pi_{g^{(\ell)}}$ by $\pi_{\tilde{g}^{(\ell)}}$ is done by normalizing $\tilde{g}^{(\ell)}$ after solving the variational problem. We prove convergence bounds of our method and compare it to existing bounds for the Hellinger distance. We demonstrate the applicability of our proposed method in Bayesian inverse problems and on graphical models.

Short biography (PhD student)

I am an electrical engineer, now pursuing a PhD in applied mathematics at the Université Grenoble Alpes while working at INRIA in the AIRSEA team. In my thesis, I work on estimation of densities for applications in variational inference.

References

[1] Tiangang Cui and Sergey Dolgov. Deep composition of tensor-trains using squared inverse Rosenblatt transports. *Foundations of Computational Mathematics*, September 2021. arXiv:2007.06968 [cs, math, stat].

[2] Tiangang Cui, Sergey Dolgov, and Olivier Zahm. Scalable conditional deep inverse Rosenblatt transports using tensor trains and gradient-based dimension reduction. *Journal of Computational Physics*, 485:112103, 2023.

[3] Tiangang Cui, Sergey Dolgov, and Olivier Zahm. Self-reinforced polynomial approximation methods for concentrated probability densities, March 2023. arXiv:2303.02554 [cs, math, stat].

[4] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference, April 2021. arXiv:1912.02762 [cs, stat].

[5] Josephine Westermann and Jakob Zech. Measure transport via polynomial density surrogates, November 2023. arXiv:2311.04172 [cs, math, stat].