

Chair in applied mathematics OQUAIDO Activity report

Olivier Roustant, Rodolphe Le Riche, Josselin Garnier, David Ginsbourger, Yves Deville, Céline Helbert, Luc Pronzato, Clémentine Prieur, Fabrice Gamboa, François Bachoc, et al.

▶ **To cite this version:**

Olivier Roustant, Rodolphe Le Riche, Josselin Garnier, David Ginsbourger, Yves Deville, et al.. Chair in applied mathematics OQUAIDO Activity report. [Research Report] Mines Saint-Etienne; Ecole Centrale Lyon; BRGM (Bureau de recherches géologiques et minières); CEA; IFP Energies Nouvelles; Institut de Radioprotection et de Sûreté Nucléaire; Safran Tech; Storengy; CNRS; Université Grenoble - Alpes; Université Nice - Sophia Antipolis; Université Toulouse 3 (Paul Sabatier). 2021. hal-03217277v2

HAL Id: hal-03217277
https://hal.science/hal-03217277v2

Submitted on 7 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Chair in applied mathematics OQUAIDO
Activity report

April 2021

The research Chair OQUAIDO – for "Optimisation et QUAntification d'Incertitudes pour les Données Onéreuses" in French – gathers academic and technological research partners to work on statistical learning problems involving scarce and error-prone data. This Chair, created in January 2016 for a period of 5 years, is the continuation of the projects DICE and ReDICE which respectively covered the periods 2006-2009 and 2011-2015. It is a joint effort between :

- Mines Saint-Étienne (that hosts the Chair), École Centrale de Lyon, CNRS, Univ. Grenoble Alpes, Univ. de Nice, Univ. de Toulouse III, as academic partners ;
- BRGM, CEA, IFPEN, IRSN, Safran, Storengy, as technological research partners ;
- Y. Deville (AlpeStat), J. Garnier (École Polytechnique), D. Ginsbourger (Idiap and Univ. of Bern), L. Polès (XtraFormation), as experts, and L. Carraro, as advisor.

Website : oquaido.emse.fr

With the current boom in data sources, Artificial Intelligence (AI) experiences a spectacular revival with implications in all domains. AI actually encompasses diverse methodologies and while those devoted to big data are the most visible to the general public, those that tackle *small data* are of utmost importance : many experiments, either real or coming from modeling through intensive computing, can be repeated only a limited number of times, yielding scarce data ; these data are affected by measurement or calculation errors ; and there exists additional qualitative or quantitative information conveyed by experts.

The OQUAIDO research chair tackles problems where small data is described by statistical models that, in turn, serve to characterize uncertainties, calibrate computer codes and search for optimal configurations. Many of the investigated approaches rely on *Gaussian processes* and confront mathematical challenges such as high dimension (even functional inputs / outputs), mixed continuous and categorical inputs, specific constraints and medium data.

This activity report highlights noticeable contributions of OQUAIDO, provides bibliography indicators and summarizes the events that have marked the research Chair life.

Authors and acknowledgement

The writing of this report has been directed by Olivier Roustant and Rodolphe Le Riche (EMSE), the current project leaders of the OQUAIDO Chair. Inputs were provided by the scientific experts, Josselin Garnier (École Polytechnique), David Ginsbourger (Idiap and Univ. of Bern), Yves Deville (AlpeStat), as well as the members of the steering committee : Céline Helbert (EC Lyon), Luc Pronzato (CNRS / UNICE), Clémentine Prieur (UGA), Fabrice Gamboa and François Bachoc (UPS), Jérémy Rohmer (BRGM), Guillaume Perrin (CEA DAM), Amandine Marrel and Guillaume Damblin (CEA DEN), Alain Glière (CEA DRT), Delphine Sinoquet (IFPEN), Yann Richet (IRSN), Sébastien Da Veiga (SAFRAN), Frédéric Huguet (Storengy).

We are grateful to all the participants. In particular, we thank Nicolas Durrande (EMSE) for leading the project with us for the first two years, Ludovic Polès (Xtra formation) for sharing his skills in project management. Last but not least, we acknowledge the mentorship of Laurent Carraro, who initiated the original type of research projects about computer experiments that have led to OQUAIDO.

Table of contents

1 Scientific program and resources **4**
- 1.1 Case studies . . . 4
- 1.2 Post-doc, PhD and master thesis funded by the Chair . . . 4

2 Chair life, at a glance **5**

3 Scientific production **6**

4 Examples of accomplishments **7**
- 4.1 Categorical inputs . . . 7
  - 4.1.1 Metamodeling in presence of categorical inputs . . . 7
  - 4.1.2 Optimization with mixed continuous and discrete variables . . . 8
- 4.2 Stochastic codes . . . 9
  - 4.2.1 Constrained optimization in the presence of uncertainties . . . 9
- 4.3 Functional inputs/outputs, high number of inputs . . . 10
  - 4.3.1 Low-cost screening of non-influential inputs . . . 10
  - 4.3.2 Inverse problem with functional inputs . . . 11
- 4.4 Specific constraints . . . 12
  - 4.4.1 Metamodeling with monotonicity / inequality constraints . . . 12
- 4.5 High number of data . . . 13
  - 4.5.1 Metamodeling with large data sets . . . 13
- 4.6 The kergp software : a laboratory to build kernels . . . 14
- 4.7 Other topics . . . 15
  - 4.7.1 Optimisation / inversion guided by a mixture of metamodels . . . 15
  - 4.7.2 Improving prediction accuracy with designs based on mutual information . . . 15

5 Lessons learnt and perspectives **16**
- 5.1 An efficient model for collaborative research . . . 16
- 5.2 Related projects on statistical learning to leverage simulation . . . 16
- 5.3 What comes next ? . . . 17

A Details of the scientific production **18**
- A.1 Software, notebooks and vignettes . . . 18
- A.2 PhD thesis . . . 18
- A.3 Publications in journals . . . 18
- A.4 Preprints . . . 19
- A.5 Conference proceedings . . . 20
- A.6 Interactions with other PhD thesis / post-docs . . . 20
- A.7 Invited talks and courses . . . 21
  - A.7.1 Courses . . . 21
  - A.7.2 Selected invited talks . . . 21

1 Scientific program and resources

The research program is at the intersection between the operational goals and scientific limitations described in Table 1. Four case studies have been proposed by partners from technological research. In addition to partners manpower, the Chair has funded 2 post-docs, 3.2 PhDs thesis, and 1 master thesis.

\ **Application** **Framework** \	Optimization	Inversion / Calibration	Uncertainty Quantification	Modeling and other
Categorical inputs	PhD 4	Case 2 - Case 2' PhD 4		Post-doc 1
Stochastic codes	Case 4 - Post-doc 2		Case 4 - Post-doc 2	
Functional inputs/outputs High nb of inputs	Case 3 - PhD 1	Case 1 - Case 3 PhD 1, MSc	PhD 1	
Specific constraints			PhD 2	PhD 2
High nb of data				
Other topics				PhD 3

TABLE 1 – Scientific program, case studies (Case) and extra manpower : post-doc, PhD thesis (PhD), master thesis (MSc).

1.1 Case studies

Case 1 BRGM, EC Lyon, Univ. Nice. *Inversion of hydrodynamical and temporal offshore conditions leading to marine submersion of the coast.*

Case 2, 2' CEA-DAM, Mines Saint-Étienne. *Metamodeling of computer codes with mixed inputs and inverse problems. Application to radionucleide quantification by Gamma spectometry.*

Case 3 IRSN, Univ. Toulouse. *Inversion of the nuclear criticity coefficient with functional inputs.*

Case 4 Safran, EC Lyon. *Robust optimization. Application on the rotor37.*

1.2 Post-doc, PhD and master thesis funded by the Chair

Post-doc 1 E. Padonou (2016 - 2017). Post-doc on *Metamodeling in presence of categorical inputs.*

Post-doc 2 R. El Amri & J. Pelamatti (2019 - 2020). Post-doc on *Constrained optimization under uncertainty.*

PhD 1 R. El Amri (2016 - 2019). PhD on *Inversion with functional inputs / outputs under uncertainty.*
Supervision team : C. Prieur (Univ. Grenoble Alpes), C. Helbert (EC Lyon), D. Sinoquet (IFPEN), O. Lepreux (IFPEN) and M. Munoz Zuniga (IFPEN).

PhD 2 A.-F. López Lopéra (2016 - 2019). PhD on *Metamodeling under inequality constraints.*
Supervision team : O. Roustant (Mines St-Étienne), F. Bachoc (Univ. Toulouse), N. Durrande (Prowler.io).

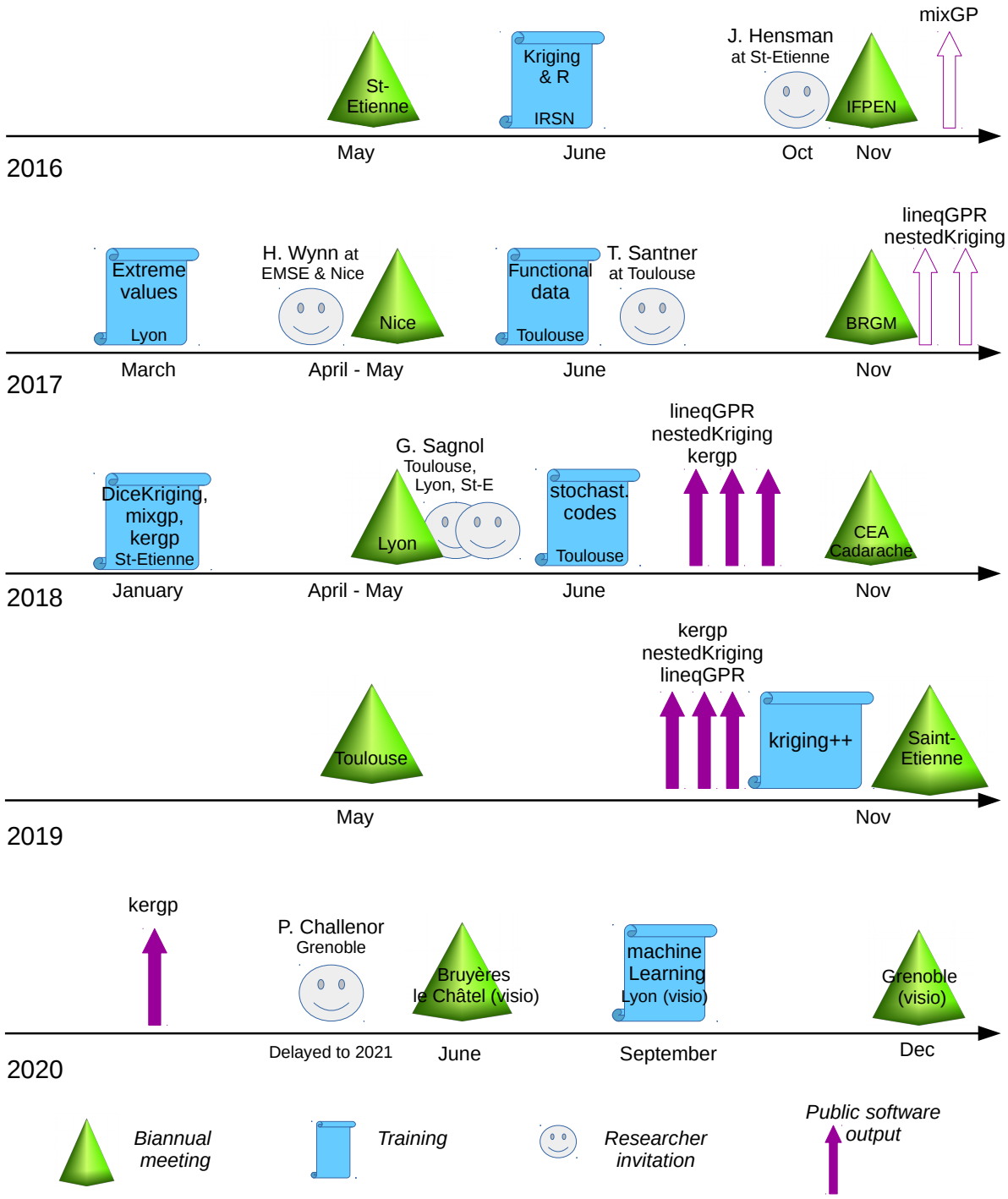
PhD 3 M. Abtini (6 months funding in 2018). PhD on *Sequential designs for Kriging.*
Supervision team : L. Pronzato (Univ. Nice), M.-J. Rendas (Univ. Nice), C. Helbert (EC Lyon).

PhD 4 J. Cuesta-Ramirez (2018 -). PhD on *Optimization with mixed continuous and discrete inputs.*
Supervision team : O. Roustant (IMT), A. Glière (CEA), G. Perrin (CEA), R. Le Riche (Mines St-Étienne).

MSc F. Allaire (2017). MSc on *Support vector regression with an adaptive criterion.*

2 Chair life, at a glance

The Chair life has been punctuated by biannual internal meetings and training sessions, in which participants discussed problems, showed solutions or learnt advanced methods. It has been also enriched by interactions with internationally recognized researchers : J. Hensman (Lancaster University, UK), H. Wynn (London School of Economics, UK), T. Santner (Ohio State University, US), G. Sagnol (Technische Universität Berlin).



3 Scientific production

Table 1 summarizes the publications that acknowledge the Chair for partial funding or privileged links (see the full list in the appendix A). In addition to standard academic outputs, such as publications in journals, conference proceedings and preprints, the Chair has provided software and notebooks, in order to enhance the usage of the methodological findings in a daily practice. The table shows the efforts spent on the modeling issue (last column), on which depends the application problems (other columns). All kinds of applications (columns) and scientific challenges (lines) have been addressed. We have nearly reached the aim to check all the boxes at the intersection of a scientific challenge and an application problem. Figure 1 shows the cross contributions of partners in the scientific production. We can see that most of the outputs involve at least two partners. The visible clusters mainly correspond to the assigned resources (see Table 1) of OQUAIDO, which has structured the collaborations between partners.

\ **Application** **Scient. challenge** \	Optimization	Inversion / Calibration	Uncertainty Quantification	Modeling
Categorical inputs	J_{13}, P_5			S_1, $N_{1,.}$, J_{12}, P_8
Stochastic codes	P_3, P_4	J_{15}		
Functional Inputs/Outputs		D_1, J_7, J_{15}, P_6	D_1, J_{16}	P_9
High nb of inputs	J_{10}	J_{10}	J_2, J_5, J_{14}	J_{10}
Specific constraints			D_2, P_2	D_2, S_2, $N_{2,.}$, J_6, J_8, C_2, C_3
High nb of data				S_3, N_3, S_4, J_3, J_4, P_1
Other topics	J_1	J_1, J_9		D_3, S_1, S_5, J_{11}, C_1, P_7

TABLE 2 – Cross classification scientific production (see details on the next pages). The capital letter meaning is as follows. S : software ; N : notebooks ; D : PhD thesis ; J : publication in journals ; P : preprint ; C : conference proceeding.

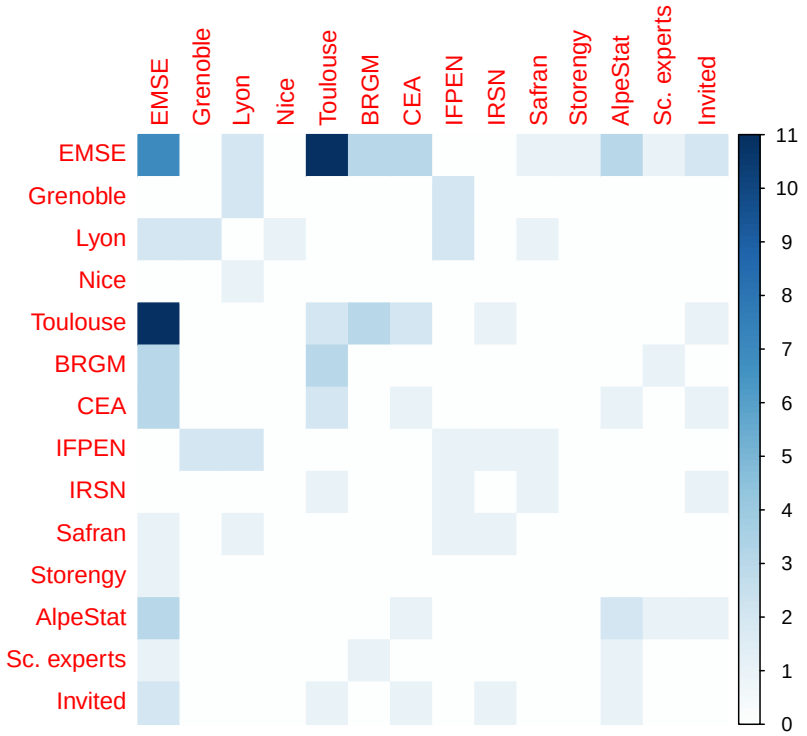
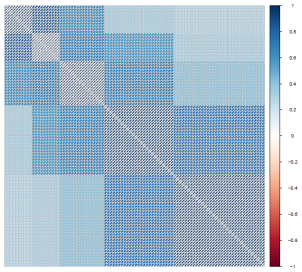


FIGURE 1 – Table of contributions : number of publications (of all type) per partner and pairs of partners. The partners are ranked by category (academic, technological research, experts, invited) and in alphabetical order inside each category.

4 Examples of accomplishments

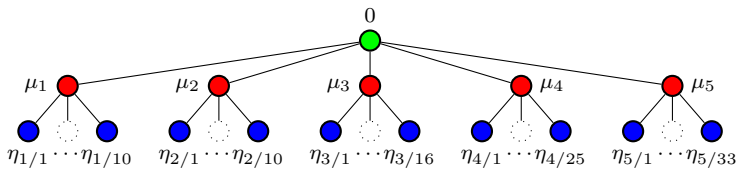
4.1 Categorical inputs

4.1.1 Metamodeling in presence of categorical inputs



Often, decision problems have some variables that are categorical (or qualitative) while other variables are quantitative. GP modeling techniques can be adapted to this context by combining continuous covariance kernels with covariance matrices. In an application to nuclear waste engineering from CEA, we had to deal with a demanding case where one categorical input, the atomic number, has a large number of levels i.e. 94. We proposed a hierarchical Gaussian model exploiting a partition in 5 groups provided by experts.

FIGURE 2 – Covariance matrix for the waste problem (top), and associated group / level tree structure (right).



GP modeling was also successfully applied in earth sciences. We have considered the problem of predicting the wave height generated by a cyclone, depending in particular on the cyclone track, viewed as a categorical input (Figure 3). Using the dataset generated within the ANR-funded spicy project (http://spicy.brgm.fr/fr), we have applied several GP models for continuous and categorical inputs developed in OQUAIDO, based on various kernel assumptions : compound symmetry, hierarchical and ordinal. These GP models outperform other off-the-self algorithms such as random forest in terms of prediction accuracy. In addition, they reveal the correlation of the output when one input is modified, and show here that the correlation varies almost monotonically with the angle of the cyclone track.

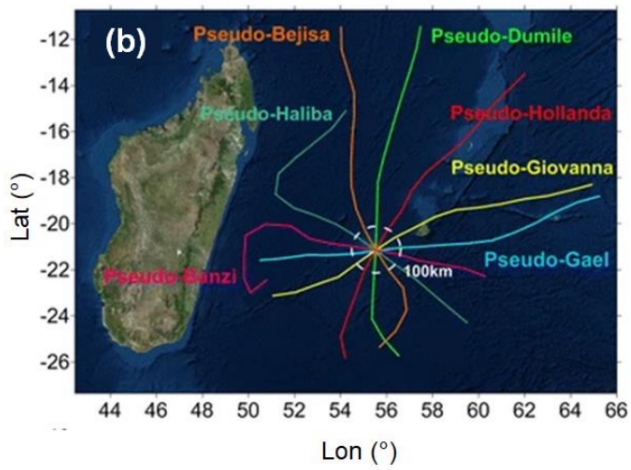


FIGURE 3 – Cyclone tracks used for modelling the waves at Saint Suzanne city (Reunion island).

Case study 2 – Publication : J_{12}, P_8 *– Software :* S_1, $N_{1,.}$.

4.1.2 Optimization with mixed continuous and discrete variables

Optimization problems with mixed – continuous and discrete – input variables are the most general ones in practice. For example in structures, continuous variables are dimensions while the discrete variables are materials ; in neural networks, the discrete variables describe the architecture of the network (number and types of neurons) and the continuous variables are the weights. In nonlinear problems, the presence of discrete variables is a source of complexity when the size of the discrete space grows.

Yet, based on physical interpretations, the existence of a smaller number of continuous latent variables can be assumed. In structures for example, the latent variables relate to some macroscopic mechanical properties such as flexural stiffness. The search for low-dimensional latent variables can been embedded in the definition of Gaussian processes. A mapping from discrete to latent variables is learned from the data points by maximum likelihood estimation. Then, an optimization procedure is defined in the space of continuous variables augmented by the continuous latent variables. The link between the latent and the discrete variables takes the form of optimization constraints, which have been handled by augmented Lagrangians. The method has been applied to synthetic functions plus the test case of the design of a light filter, cf. Figure 4.

In parallel, we have pursued the other track where the relaxation of the discrete variables is avoided. An algorithm was devised where a Gaussian process is created over the mixed space and guides a Bayesian optimization algorithm (EGO) generalized for such spaces.

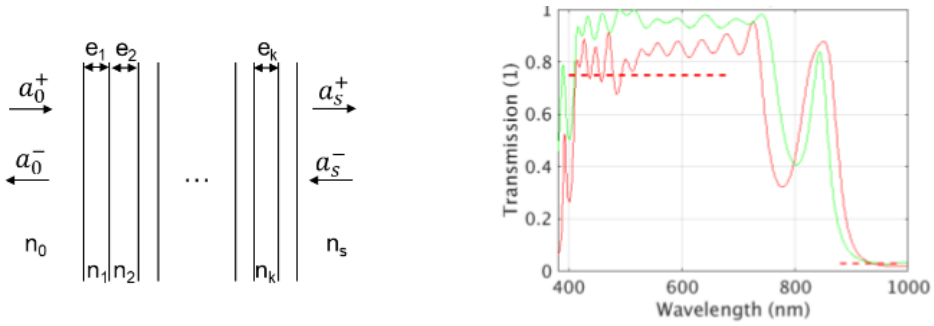


FIGURE 4 – Design of a multilayer filter for light radiations. The design variables are the materials of the layers (labelled 'n'), which are categorical, and the layers' thicknesses (labelled 'e'), which are continuous. The objective function is a distance to a target of the transmission spectrum.

PhD 4 – Publications : J_{13}, P_5.

4.2 Stochastic codes

4.2.1 Constrained optimization in the presence of uncertainties

Engineering and more generally the sciences that draw upon numerical simulation often need to solve optimization problems where evaluating a solution is expensive, where there are optimization constraints, and some of the input parameters are uncertain. During the OQUAIDO project, a methodology to tackle these problems has been proposed. It starts from the assumption that uncertain parameters have a known distribution and can be chosen as inputs to the numerical simulation. This makes it possible to build a statistical model (a kriging model) in the joined space of controlled variables (the x's) $\times$ uncertain parameters (the u's) and define optimal strategies to choose the (x, u) iteratively.

More precisely, we have addressed the case of *chance constraints*, i.e., constraints that need to be satisfied with a given confidence. We have proposed a two-step algorithm for first choosing the controlled variables x, then a sample of the uncertain variables u. The first step is a maximization of expected feasible improvement, the second a one-step-ahead variance reduction. We have shown how the variance reduction criterion can be approximated for better computing efficiency. Our statistical models have been improved to capture correlated constraints through an output-as-input strategy. Finally, our iterative procedure is particularly well adapted to expensive simulations/experiments because we have been able not only to choose the inputs to the simulations, but also to decide which part of the simulation/experiments (as a constraint) is the most relevant to be invoked.

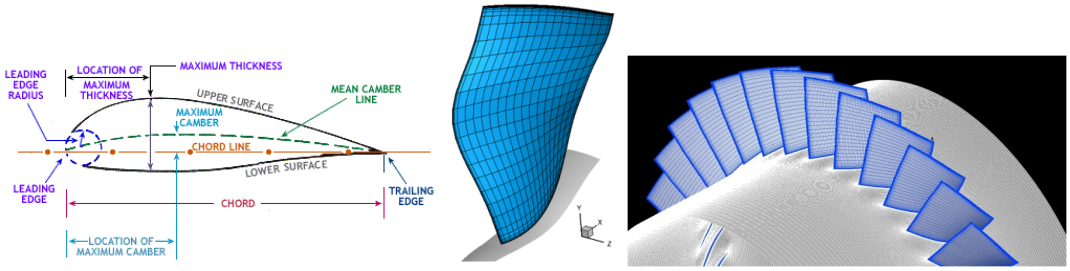


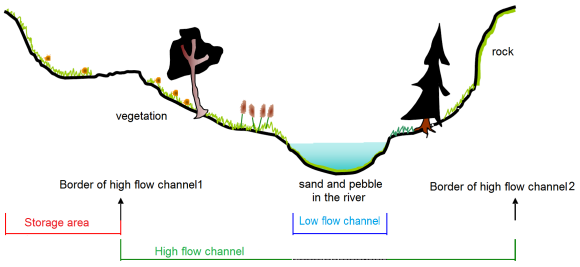
FIGURE 5 – The rotor test case : the blades are described by 4 cross-sections for a total of 20 design parameters. There are uncertainties about the manufacturing (rugosity, tip gap), the inflow properties (pressure, temperature and azimuthal momentum) and operational conditions (flow rate, rotation speed). These 7 uncertain parameters affect 5 constraints about the inlet and outlet relative flow angles, the flow speed reduction, excessive loading and the Mach number of the blade tips. The objective function is the polytropic (compressor) efficiency.

The approach has been applied to the *design of rotor blades* as proposed by Safran Tech and illustrated in Figure 5.

Post-doc 2 – Publications : P_3,P_4.

4.3 Functional inputs/outputs, high number of inputs

4.3.1 Low-cost screening of non-influential inputs



We have improved a screening method for finding the non-influential inputs of a complex computer code using its derivatives. It is based on the search of optimal Poincaré constants in functional inequalities. The surrounding illustrations describe an application to a flooding problem investigated with the Mascaret software (a code which solves the Saint Venant equations).

Publication : J_2, J_{14}.

Software : update of the R package `sensitivity`.

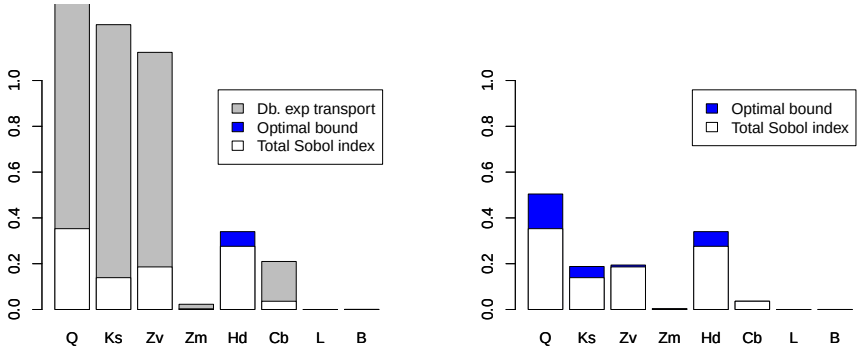
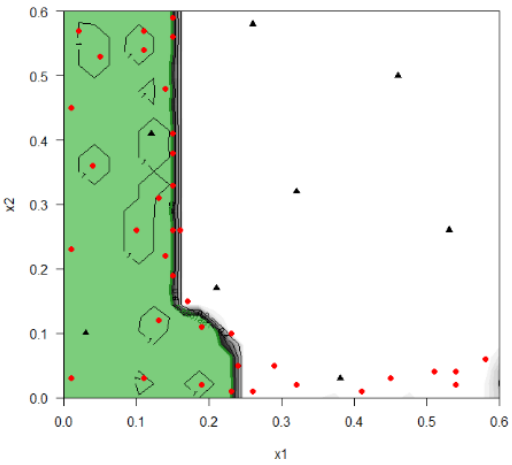


FIGURE 6 – Sobol indices (white bars) and cheap-to-evaluate upper bounds (grey and blue bars) for the flooding problem. The new technique (right) detects the 4 non-influential variables Z_m, C_b, L, B and gives sharp upper bounds of Sobol total sensitivity indices, improving on the standard technique (left).

4.3.2 Inverse problem with functional inputs

Given a complex computer code f, the problem of finding the set $\Gamma_f = \{x,\ f(x) \leq c\}$ is a frequent real-life inverse problem. When x is a vector of continuous inputs, efficient algorithms exist, which add points sequentially in order to best decrease the uncertainty of the random set Γ_Y obtained by replacing f by a GP metamodel Y. In an automotive test case from IFPEN, we were faced with two new issues : there are additional input time curves V, and these functional inputs are not controllable.
We proposed a two-stage robust algorithm which fixes the problem and estimates $\Gamma_{f,V} = \{x,\ \mathbb{E}_V(f(x,V)) \leq c\}$.
PhD 1 – Publication : D_1, J_7.



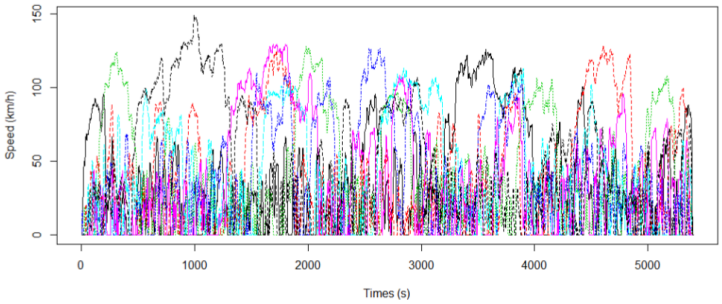


FIGURE 7 – Automotive test case : f simulates the pollutant concentration and V represent driving cycles. Left : examples of driving cycles. Top : estimation of $\Gamma_{f,V}$ (green area), after 37 iterations. Black triangles : initial data ; Red points : added data.

4.4 Specific constraints

4.4.1 Metamodeling with monotonicity / inequality constraints

Prediction and uncertainty of a statistical model can be drastically improved by accounting for inequality constraints, such as box constraints, monotonicity, convexity, leading to impressive reduction of the computational budget. A finite elements model with Gaussian random coefficients can be used for that goal. That model has been developed to deal with several inequality constraints at a time, motivated by applications from BRGM and IRSN (Fig. 8). It has also been scaled up for higher dimensions. In particular, we introduced the MaxMod algorithm that sequentially selects the active variables, which makes the method applicable in dimension 20. We proved the convergence of that dynamical algorithm to the solution of the spline problem under inequality constraints, extending a known static result.
Finally, we studied the properties of the parameter estimators under inequality constraints, and obtained asymptotic distributions. These results confirm the intuition that when the number of observations tends to infinity, the constraints are automatically learnt, and can be neglected in hyperparameter estimation. In practice, even for small sample sizes, the standard unconstrained likelihood estimator seems to achieve a good tradeoff between estimation accuracy and computational tractability.
PhD 2 – Publication : D_2, J_6, J_8, C_2, C_3, P_2 *– Software :* S_2.

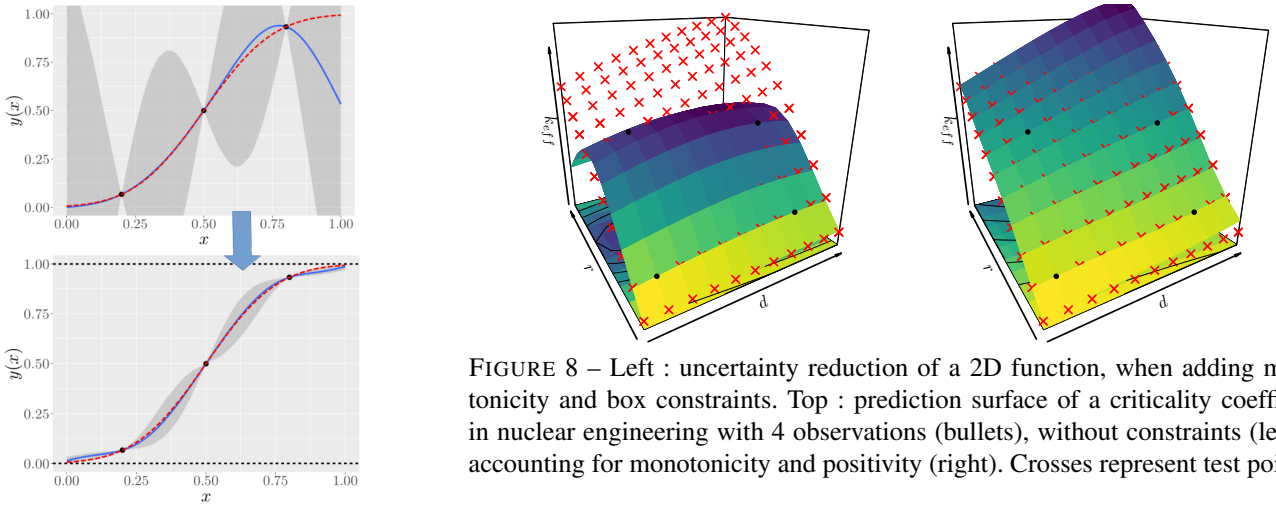


FIGURE 8 – Left : uncertainty reduction of a 2D function, when adding monotonicity and box constraints. Top : prediction surface of a criticality coefficient in nuclear engineering with 4 observations (bullets), without constraints (left) or accounting for monotonicity and positivity (right). Crosses represent test points.

4.5 High number of data

4.5.1 Metamodeling with large data sets

Gaussian Processes are popular for interpolation and uncertainty handling. Some people claim that they are not adapted for more than one thousands of data. We developed a method to deal with more than *hundred thousands* of data points : by linearly combining predictions on data subsets, one can build a useful approximation (a Gaussian Process that is interpolating, with known variance and proven consistency). The method has been implemented in the *nestedKriging R* package.

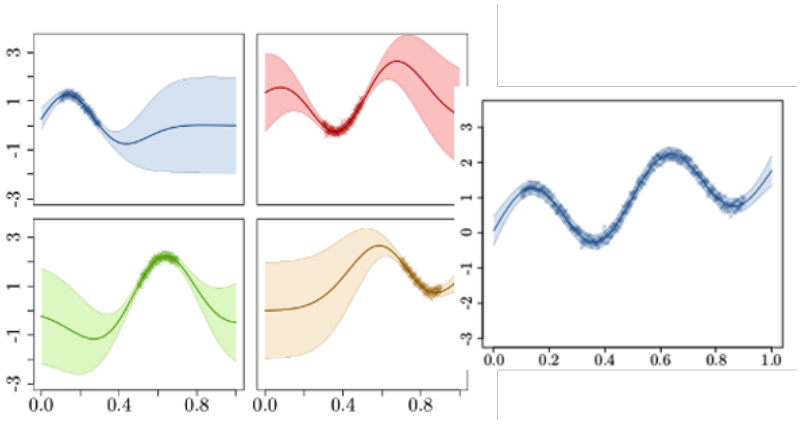


FIGURE 9 – Left : data is split in four subsets / submodels.
Right : aggregation by best pointwise linear combination of the sub-model predictions.

Furthermore, *nestedKriging* and other statistical learning methods were compared on high dimensional data of CO2 reservoirs provided by Storengy.

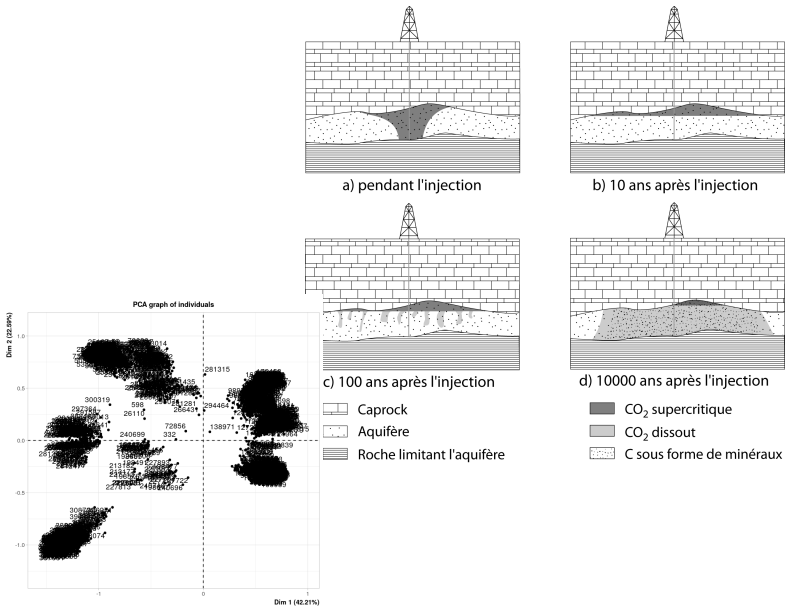


FIGURE 10 – Storengy data about CO2 storage : about 400,000 points, 34 dimensions describing reservoir structure, 4 criteria related to gas and water flows. Left : graph of individuals in the first two principal directions. Right : sketch of a reservoir.

Publications : J_4, P_1 *– Software :* S_3,N_3.

4.6 The kergp software : a laboratory to build kernels

The *kergp* package implements a range of solutions that may fit the needs of any user. One can find general classes of kernels for continuous inputs (tensor product, tensor sum, ANOVA, radial) as well as for discrete ordinal inputs (warping-based kernels) and discrete categorical inputs (compound symmetry, low rank, group kernels, ...). Furthermore, *kergp* allows to code from scratch customized kernels, with several mechanisms corresponding to various coding skill levels (from R beginners to C++ experts), requiring or not the gradient information. *kergp* also allows to *build new kernels from old ones*, thanks to a formula mechanism. For instance, once kernels for continuous inputs (`kCont`) and for discrete inputs (`kOrd`, `kCat`) have been defined, a tensor-product kernel for mixed continuous and discrete inputs can be coded by :

$$\sim \texttt{kCont}() * \texttt{kOrd}() * \texttt{kCat}()$$

After building a kernel, *kergp* enjoys all the main expected functionalities for Gaussian processes : hyperparameter estimation, validation, prediction, simulation.

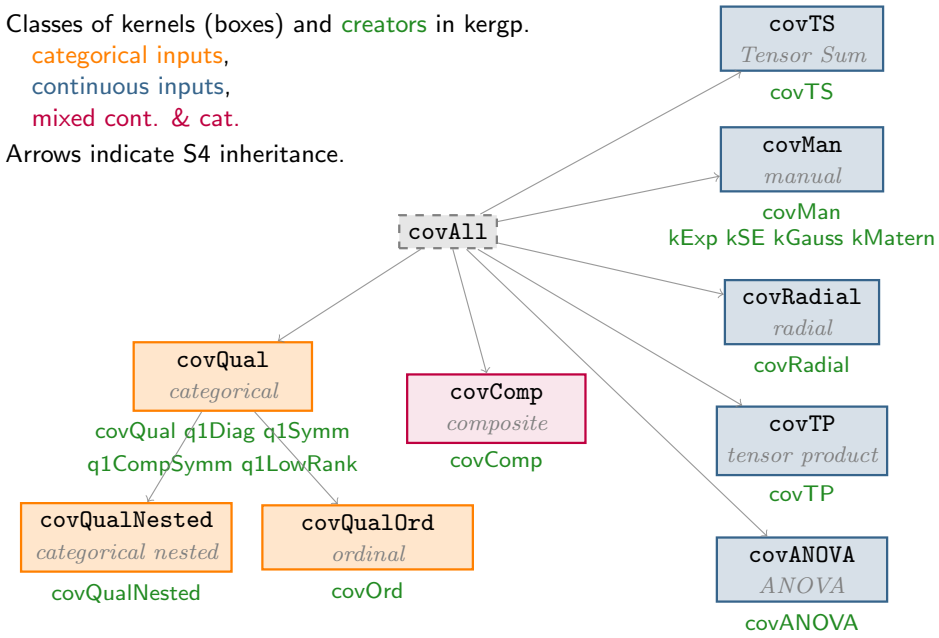


FIGURE 11 – Main classes of kernels in *kergp*

4.7 Other topics

4.7.1 Optimisation / inversion guided by a mixture of metamodels

A frequent question for practitioners using metamodel-based optimization (or inversion) problems is the choice of a metamodel type and parameters among the wide list of possibilities. Rather than choosing, we suggest to mix them in order to benefit from the advantages of all of them. This has been made possible by extending the notion of local uncertainty to a general metamodel called Universal Prediction distribution.
See publication : J_1.

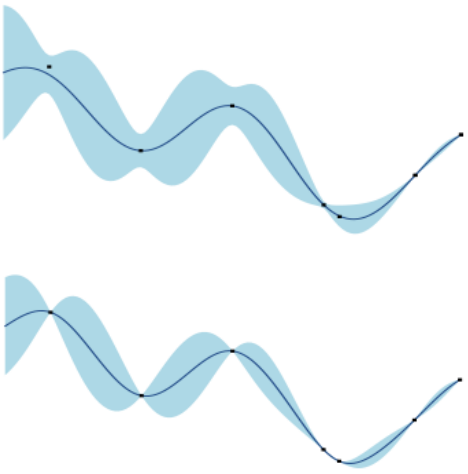
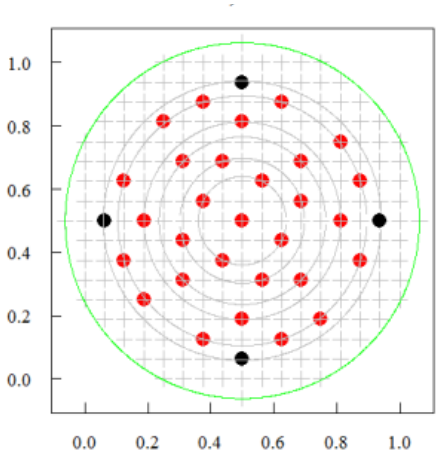


FIGURE 12 – UP distribution in action : for a SVM approximator (top) and a GP interpolator (down). The blue shaded areas are uncertainty regions.

4.7.2 Improving prediction accuracy with designs based on mutual information



While a common practice in computer experiments is to construct an initial space-filling design, a pure distance-based criterion may not guarantee a good prediction accuracy. To reach that aim, we have considered the model-based mutual information (MI) criterion, in the frame of Gaussian processes. We have developed economical sequential strategies to build MI-optimal designs.
See publication : D_3, P_7.

FIGURE 13 – MI sequential design for an isotropic Gaussian process on a circular domain. The four black points are the new added points.

5 Lessons learnt and perspectives

5.1 An efficient model for collaborative research

OQUAIDO is a consortium where academic and technological partners join their efforts to tackle a well-defined family of mathematical challenges that are shared across various applications. The collaboration model has evolved from that of the DICE and reDICE projects. It is based on the following principles :

- The partners develop **theoretical methodologies guided by industrial questions**. Test cases are addressed, with TRL's between 1 and 2. The industrial test cases are mathematically formalized. The work is grounded in theory while high quality **prototype software is produced** with the help of professional developers.
- **Teaching and knowledge sharing** through meetings and training sessions (about 20 participants per session, cf. figure in Section 2 for a detailed calendar) is a constitutive part of the project.
- **The consortium is diverse and has a medium size** : of the order of 6 academic and 6 industrial partners. This is more than typical nationally funded projects or research chairs. As a result, the project also acts as a network builder, but it is sufficiently small for people to meet and interact.
- **The project is funded by the non-academic partners** who pay a relatively low entry cost. No public funding is necessary. While this mechanism limits staff resources, it also drastically reduces paperwork, uncertainties about project approval, and every partner feels more strongly tied to each other as resources need to be shared. It often occurs that smaller groups of partners start complementary projects around a specific task and seek complementary funding (e.g., public).
- **Decisions, in particular research directions, training topics and invitations of researchers are discussed as a college**.
- **Research results are publically released**, in articles or in opensource softwares.

5.2 Related projects on statistical learning to leverage simulation

Numerical simulation and data acquisition are a growing part of research and product development. Statistical models, calibration, uncertainty propagation and optimization leverage their usefulness. For this reason, while OQUAIDO is now finished, it has and will contribute to other projects whose partners and research directions overlap with the Chair :

- The ANR SAMOURAI (Simulation Analytics and Meta-model-based solutions for Optimization, Uncertainty and Reliability AnalysIs) project will start in March 2021 with IFPEN, EDF, Safran Tech, CEA, Centrale Supélec, Mines St-Étienne, Polytechnic Montreal as partners. The principal investigator is IFPEN (Delphine Sinoquet).
- OQUAIDO has played a key role when starting complementary PhDs and post-doctorates (cf. "Interactions with other PhD thesis / post-docs" at the previous page).
- The *libKriging* platform is an effort to carry some of the R toolboxes developed during OQUAIDO, DICE and reDICE, to more efficient implementations (in C++), satisfying industrial quality standards, and offering a larger compatibility with other languages (Python, matlab, octave). The main contributors to *libKriging* come from IRSN and Haveneer and should be complemented by developers from AlpeStat, IFPEN, Safran Tech, Mines St-Étienne and others.

5.3 What comes next ?

Last but not least, OQUAIDO has a follow-up project, CIROQUO (which stands for "Consortium Industrie Recherche pour l'Optimisation et la QUantification d'incertitude pour les données Onéreuses" in French). CIROQUO will start in 2021 and builds on the experience gained during OQUAIDO, both from a scientific and from a coordination point of view. Current partners of CIROQUO are Centrale Lyon, IFPEN, Mines Saint-Étienne, CNRS, Université Côte d'Azur, Université de Toulouse, INRIA, IRSN, BRGM, CEA and Storengy with plans to broaden the consortium in 2021. It is led by Centrale Lyon (Céline Helbert and Christophette Blanchet-Scalliet) and IFPEN (Delphine Sinoquet).

We are convinced that projects like OQUAIDO have a balanced and efficient contribution to sciences and technologies. It has been a pleasure to share the lessons learned from it. We wish a lot of success to CIROQUO.

A Details of the scientific production

A.1 Software, notebooks and vignettes

S_1 *kergp* : kernel laboratory. This package, created during the ReDICE consortium, has been enriched with new functionalities : categorical variables, radial kernels, optimizer choices, etc.

$N_{1,a}$ Using the R package *kergp* : Mauna Loa CO_2 data example (2019), Y. Deville.

$N_{1,b}$ Using the R package *kergp* : group kernels (2019), O. Roustant.

$N_{1,c}$ Using the R package *kergp* : Ordinal kernels on the beam problem (2019), O. Roustant and Y. Deville.

$N_{1,d}$ Analysis categorical inputs for cyclone-induced wave modeling (2020), J. Rohmer and O. Roustant.

S_2 *lineqGPR* : Gaussian process regression models with linear inequality constraints.

$N_{2,a}$ *lineqGPR* instructions manual (2018). A.F. López-Lopera.

$N_{2,b}$ *lineqGPR* in action on a flooding problem (2020). A.F. López-Lopera.

S_3 *nestedKriging* : nested Kriging models for large datasets.

N_3 Statistical modeling of reservoir data : an empirical study (2020), H. Devaine, R. Le Riche, D. Gaudrie, D. Rullière and F. Huguet.

S_4 *specgp* : construction of kernels by the spectral approach, suitable e.g. for large datasets.

S_5 *libKriging* : this is an ongoing software project, to enhance an industrial usage of OQUAIDO results. *libKriging* will include fast and portable implementations for GP modeling, with a wide test coverage.

Remark. Another package, called *mixgp*, dedicated to Kriging models with both discrete and continuous input variables, has been developed in the first years of the Chair. It is now included in *kergp*.

A.2 PhD thesis

D_1 R. El Amri, *Analyse d'incertitudes et de robustesse pour les modèles à entrées et sorties fonctionnelles*, Phd thesis, Université Grenoble Alpes, April 2019.

D_2 A.F. López-Lopera, *Gaussian Process Modeling under Inequality Constraints*, PhD thesis, Université de Lyon, September 2019.

D_3 M. Abtini, *Plans prédictifs à taille fixe et séquentiels pour le krigeage*, PhD thesis, École Centrale Lyon, August 2018.

A.3 Publications in journals

J_1 Universal Prediction Distribution for Surrogate Models, M. Ben Salem, O. Roustant, F. Gamboa, and L. Tomaso (2017), *SIAM/ASA Journal on Uncertainty Quantification*, **5** (1), 1086-1109.

J_2 Poincaré inequalities on intervals - application to sensitivity analysis, O. Roustant, F. Barthe and B. Iooss (2017), *Electronic Journal of Statistics*, **11** (2), 3081-3119.

J_3 Variational Fourier Features for Gaussian Processes J. Hensman, N. Durrande and A. Solin (2018), *Journal of Machine Learning Research*, **8**, 1-52.

J_4 Nested Kriging predictions for datasets with a large number of observations, D. Rullière, N. Durrande, F. Bachoc and C. Chevalier (2018), *Statistics and Computing*, **28** (4), 849-867.

J_5 Sensitivity Analysis Based on Cramér von Mises Distance, F. Gamboa, T. Klein, and A. Lagnoux (2018), *SIAM/ASA Journal on Uncertainty Quantification*, **6** (2), 522-548.

J_6 Finite-dimensional Gaussian approximation with linear inequality constraints, A.F. López-Lopera, F. Bachoc, N. Durrande and O. Roustant (2018), *SIAM/ASA Journal on Uncertainty Quantification*, **6** (3), 1224–1255.

J_7 Data-driven stochastic inversion under functional uncertainties, M.R. El Amri, C. Helbert, O. Lepreux, M. Munoz Zuniga, C. Prieur and D. Sinoquet (2020), *Statistics and Computing*, **30** (3), 525-541.

J_8 Maximum likelihood estimation for Gaussian processes under inequality constraints, F. Bachoc, A. Lagnoux and A.F. López-Lopera (2019), *Electronic Journal of Statistics*, **13** (2), 2921-2969.

J_9 Profile extrema for visualizing and quantifying uncertainties on excursion regions. Application to coastal flooding. D. Azzimonti, D. Ginsbourger, J. Rohmer and D. Idier (2019), *Technometrics*, **61** (4), 474-493.

J_{10} Sequential dimension reduction for learning features of expensive black-box functions, M. Ben Salem, F. Bachoc, O. Roustant, F. Gamboa and L. Tomaso (2019), *SIAM/ASA Journal on Uncertainty Quantification*, **7** (4), 1369-1397.

J_{11} Karhunen-Loève decomposition of Gaussian measures on Banach spaces, X. Bay and J.C. Croix (2019), *Probability and Mathematical Statistics*, **39** (2), 279-297.

J_{12} Group kernels for Gaussian process metamodels with categorical inputs, O. Roustant, E. Padonou, Y. Deville, A. Clément, G. Perrin, J. Giorla and H. Wynn (2020), *SIAM/ASA Journal on Uncertainty Quantification*, **8** (2), 775-806.

J_{13} Global optimization for mixed categorical-continuous variables based on Gaussian process models with a randomized categorical space exploration step, M. Munoz Zuniga and D. Sinoquet (2020), *INFOR Journal*, **58**, 310-341.

J_{14} Parseval inequalities and lower bounds for variance-based sensitivity indices, O. Roustant, F. Gamboa, B. Iooss (2020), *Electronic Journal of Statistics*, **14** (1), 386-412.

J_{15} Sequential design of mixture experiments with an empirically determined input domain and an application to burn-up credit penalization of nuclear fuel rods, F. Bachoc, T. Barthe, T. Santner, Y. Richet (2021), to appear in *Nuclear Engineering and Design*, **374**.

J_{16} Functional principal component analysis for global sensitivity analysis of model with spatial output, T.V.E. Perrin, O. Roustant, J. Rohmer, O. Alata, J.P. Naulin, D. Idier, R. Pedreros, D. Moncoulon, P. Tinard, to appear in *Reliability Engineering & System Safety* (2021).

A.4 Preprints

P_1 Some properties of nested Kriging predictors, F. Bachoc, N. Durrande, D. Rullière and C. Chevalier (2017).

P_2 Sequential construction and dimension reduction of Gaussian processes under inequality constraints, F. Bachoc, A. F. López-Lopera and O. Roustant (2020).

P_3 A sampling criterion for constrained Bayesian optimization with uncertainties, R. El Amri, R. Le Riche, C. Helbert, C. Blanchet-Scalliet, S. Da Veiga (2021).

P_4 Coupling constraints in Bayesian optimization, J. Pelamatti, R. Le Riche, C. Helbert, C. Blanchet-Scalliet, to appear (2021).

P_5 Optimization in presence of categorical inputs with latent variables, J. Cuesta-Ramirez, C. Durantin, A. Glière, G. Perrin, R. Le Riche, O. Roustant, to appear (2021).

P_6 Set inversion under functional uncertainties with joint meta-models, R. El Amri, C. Helbert, M. Munoz-Zuniga, C. Prieur, D. Sinoquet (2020).

P_7 Sequential design for prediction with Gaussian process models, M. Abtini, C. Helbert, F. Musy, L. Pronzato, M.-J. Rendas (2020).

P_8 Revealing the dependence structure of scenario-like inputs in numerical environmental simulations using Gaussian Process regression, J. Rohmer, O. Roustant, S. Lecacheux, J.-C. Manceau (2020)

P_9 Multi-output Gaussian processes with functional data : a study on coastal flood hazard assessment, A. F. López-Lopera, D. Idier, J. Rohmer, F. Bachoc (2020).

A.5 Conference proceedings

C_1 Gaussian Processes For Computer Experiments, F. Bachoc, E. Contal, H. Maatouk, and D. Rullière (2017), *ESAIM Proceedings and surveys, proceedings of MAS2016 conference*, **60**, 163-179.

C_2 Gaussian Process Modulated Cox Processes under Linear Inequality Constraints, A. F. López-Lopera, S. John, and N. Durrande (2019), *PMLR :, proceedings of AISTATS19 conference*, **89**, 1997-2006.

C_3 Approximating Gaussian Process Emulators with Linear Inequality Constraints and Noisy Observations via MC and MCMC, A. F. López-Lopera, F. Bachoc, N. Durrande, J. Rohmer, D. Idier, and O. Roustant (2019), *Monte Carlo and Quasi-Monte Carlo Methods : proceedings of MCQMC18 conference*, 363-381.

A.6 Interactions with other PhD thesis / post-docs

- M. Ben Salem, PhD on *Model selection and adaptive sampling in metamodeling* (Ansys, Mines SE, IMT).
- B. Broto, PhD on *Sensitivity analysis with dependent random variables* (CEA & IMT).
- T. Browne, PhD on *Sensitivity analysis of stochastic computer codes* (EDF, Univ. Paris V).
- M.-L. Cauwet, post-doc on *Optimization in presence of categorical inputs* (Mines Saint-Étienne).
- A. Cousin, PhD on *Optimization under probabilistic constraints* (IFPEN, Ecole Polytechnique).
- J.-C. Croix, PhD on *Inverse problems in Banach spaces* (Mines Saint-Étienne).
- N. Garland, PhD on *Nested computer codes* (IRSN, Mines Saint-Étienne).
- D. Gaudrie, PhD on *Multiobjective optimization* (PSA, Mines Saint-Étienne, INRA).
- C. Haberstich, PhD on *Approximation of high-dimensional functions with tree tensor networks* (Univ. Nantes).
- B. Kerleguer, PhD on *Multifidelity models with functional input and outputs* (CEA).
- S. Marque-Pucheu, PhD on *Gaussian process regression of nested computer codes* (CEA).
- A. Meynaoui, PhD on *New dependence measures for sensitivity analysis*, (IMT).
- J. Muré, PhD on *Bayesian inference for GP metamodels* (EDF, École Polytechnique).
- T.V.E. Perrin, PhD on *Metamodeling and sensitivity analysis for models with spatial outputs* (Mines SE).
- M. Ribaud, PhD on *Metamodeling and multiobjective optimization* (École Centrale Lyon).

- R. Ravaille, PhD on *Gaussian processes for images* (Univ. St-Étienne).
- A. Spagnol, PhD on *Optimization and sensitivity analysis* (Safran Tech, Mines Saint-Étienne).
- J. Stenger, PhD on *Optimal Uncertainty Quantification of a risk measurement from a computer code*, IMT.
- L. Torossian, PhD on *Interactions between machine learning and computer experiments* (INRA, IMT).
- T. T. Tran, PhD on *Nonlinear optimization problem with mixed continuous and discrete variables* (IFPEN, Safran Tech, ENAC).

A.7 Invited talks and courses

A.7.1 Courses

- J. Hensman (Prowler.io), *Variational inference.*
- D. Ginsbourger (Idiap and Univ. of Bern), *Positive definite functions.*
- J. Garnier (École Polytechnique), *Inverse problems*.
- D. Ginsbourger (Idiap and Univ. of Bern), *Methods for uncertainty quantification / reduction on random sets.*
- G. Sagnol (TU Berlin), *Convex optimization*.
- M. Keller (EDF), *Bayesian calibration*.
- S. Puechmorel (ENAC) and A. Le Brigant (ENAC & IMT), *Centroids in non-Euclidean spaces*.
- M. Sebag (LRI & CNRS), *Causal modeling*.
- I. Redko (Univ. Jean Monnet, St-Étienne), *Transfer learning.*

A.7.2 Selected invited talks

- M. Blazère (Institut de Mathématiques de Toulouse), *Dimension reduction methods beyond PCA*.
- M. Alvarez (Univ. Tecnológica de Pereira), *Gaussian processes in Applied Neuroscience : A case study in Deep Brain Stimulation*.
- M. Filippone (Eurecom), *Unbiased computations for tractable and scalable learning of Gaussian processes*.
- H. Wynn (London School of Economics). *Tube-based bounds for Gaussian Process emulation, based on smooth polynomial methods*.
- E. Tric et S. Migeon (GeoAzur), *Risk of tsunami*.
- L. Gilquin (INSA Lyon) et C. Marteau (ICJ Lyon), *Localization and caracterization of acoustic sources*.
- G. Vial (Institut Camille Jordan), *Shape optimization : specificities and suitable numerical methods.*
- F. Ferranti (IMT Atlantique), *Uncertainty quantification for large scale systems*.
- A. Usseglio-Carleve (INRIA), *Elliptic random fields*.
- T. Espinasse (Univ. Lyon) and P. Rochet (Univ. Nantes), *Utilisation of statistics in graphs*.
- N. Durrande (Prowler.io), *Gaussian Markov random fields and sparse precision matrices*.
- D. Brockhoff (INRIA), *Multiobjective (blackbox) optimization via single-objective solvers*.

- J. Garnier (École polytechnique), *Epidemiology models and Covid-19*.
- D. Ginsbourger (Idiap and Univ. of Bern), *Modeling and optimizing set functions via RKHS embeddings*.
- D. Ginsbourger (Idiap and Univ. of Bern), *k-fold validation for Kriging*.