RESEARCH CENTRE

Bordeaux - Sud-Ouest

2020
ACTIVITY REPORT

Team
CQFD

Quality control and dynamic reliability

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions)

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Stochastic approaches

Contents

Team CQFD 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3
2.1 Presentation . . . 3

3 Research program 3
3.1 Introduction . . . 3
3.2 Main research topics . . . 3

4 Application domains 5
4.1 Dependability and safety . . . 5

5 Highlights of the year 5

6 New software and platforms 6
6.1 New software . . . 6
6.1.1 vimplclust . . . 6
6.1.2 dd-sPLS . . . 6
6.1.3 outlierSIR . . . 6

7 New results 7
7.1 Power-of-d-Choices with Memory: Fluid Limit and Optimality . . . 7
7.2 A Convex Programming Approach for Discrete-Time Markov Decision Processes under the Expected Total Reward Criterion . . . 7
7.3 On the expected total cost with unbounded returns for Markov decision processes . . . 7
7.4 A backward Itô–Ventzell formula with an application to stochastic interpolation . . . 7
7.5 A duality formula and a particle Gibbs sampler for continuous time Feynman-Kac measures on path spaces . . . 8
7.6 A second order analysis of McKean–Vlasov semigroups . . . 8
7.7 A perturbation analysis of stochastic matrix Riccati diffusions . . . 8
7.8 Applicability and Interpretability of Ward's Hierarchical Agglomerative Clustering With or Without Contiguity Constraints . . . 9
7.9 La modélisation de l'indemnisation du préjudice corporel : Un exemple de «justice quantitative» au service de l'équité . . . 9
7.10 EEG Feature Extraction Using Genetic Programming for the Classification of Mental States 9
7.11 Alternative Ways to Compare the Detendred Fluctuation Analysis and its Variants. Application to Visual Tunneling Detection . . . 10
7.12 SOAP: Semantic Outliers Automatic Preprocessing . . . 10
7.13 A Variable Chirp Rate Stepped Frequency Linear Frequency Modulation Waveform Designed to Approximate Wideband Non-Linear Radar Waveforms . . . 10
7.14 Sea Target Classification Based On An A Priori Motion Model . . . 11
7.15 Regularized Dfa To Study The Gaze Position Of An Airline Pilot . . . 11
7.16 Unlabeled Multi-Target Regression with Genetic Programming . . . 11
7.17 Reconstructing climatic modes of variability from proxy records using ClimIndRec version 1.0 11
7.18 Various Statistical Approaches to Assess and Predict Carcass and Meat Quality Traits . . . 12
7.19 Multiple-output quantile regression through optimal quantization . . . 12
7.20 BIG-SIR: a Sliced Inverse Regression approach for massive data . . . 12
7.21 A computational methodology for multi-objective optimization in precision rearing . . . 13

8 Bilateral contracts and grants with industry 13
8.1 Bilateral contracts with industry . . . 13

9 Partnerships and cooperations **14**
9.1 International initiatives . . . 14
9.1.1 Inria international partners . . . 14
9.2 National initiatives . . . 14

10 Dissemination **15**
10.1 Promoting scientific activities . . . 15
10.1.1 Scientific events: organisation . . . 15
10.1.2 Scientific events: selection . . . 15
10.1.3 Journal . . . 15
10.1.4 Leadership within the scientific community . . . 16
10.1.5 Scientific expertise . . . 16
10.1.6 Research administration . . . 16
10.2 Teaching - Supervision - Juries . . . 16

11 Scientific production **18**
11.1 Major publications . . . 18
11.2 Publications of the year . . . 18
11.3 Cited publications . . . 21

Team CQFD

Creation of the Project-Team: 2009 January 01, end of the Team: 2020 December 31

Keywords

Computer sciences and digital sciences

A3.3. – Data and knowledge analysis

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.5. – Bayesian methods

A3.4.6. – Neural networks

A3.4.7. – Kernel methods

A5.9.2. – Estimation, modeling

A5.9.6. – Optimization tools

A6.1.2. – Stochastic Modeling

A6.1.3. – Discrete Modeling (multi-agent, people centered)

A6.2.2. – Numerical probability

A6.2.3. – Probabilistic methods

A6.2.4. – Statistical methods

A6.2.6. – Optimization

A6.4.2. – Stochastic control

A6.4.6. – Optimal control

A9.2. – Machine learning

A9.6. – Decision support

Other research topics and application domains

B1.2.2. – Cognitive science

B2.2.4. – Infectious diseases, Virology

B2.6.1. – Brain imaging

B5.9. – Industrial maintenance

B6.2. – Network technologies

B6.3.3. – Network Management

B6.5. – Information systems

B9.2.3. – Video games

B9.5.2. – Mathematics

B9.5.3. – Physics

B9.5.6. – Data science

B9.11. – Risk management

B9.11.1. – Environmental risks

B9.11.2. – Financial risks

1 Team members, visitors, external collaborators

Research Scientists

- Pierre Del Moral [Inria, Senior Researcher, HDR]
- Emma Horton [Inria, Researcher, from Dec 2020]

Faculty Members

- Francois Dufour [Team leader, Institut National Polytechnique de Bordeaux, Professor, HDR]
- Marie Chavent [Univ de Bordeaux, Associate Professor, HDR]
- Alexandre Genadot [Univ de Bordeaux, Associate Professor]
- Pierrick Legrand [Univ de Bordeaux, Associate Professor, HDR]
- Jerome Saracco [Institut National Polytechnique de Bordeaux, Professor, HDR]
- Huilong Zhang [Univ de Bordeaux, Associate Professor]

Post-Doctoral Fellow

- Hadrien Lorenzo [Inria, from Sep 2020]

PhD Students

- Bastien Berthelot [Thales, CIFRE]
- Tiffany Cherchi [Thales, CIFRE]
- Alexandre Conanec [Bordeaux Sciences Agro]
- Loic Labache [CEA, CIFRE, until Oct 2020]
- Alex Mourer [Groupe SAFRAN, CIFRE]
- Camille Palmier [ONERA]
- Nathanael Randriamihamison [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement]

Interns and Apprentices

- Romain Namyst [Inria, from Jun 2020 until Aug 2020]

Administrative Assistant

- Catherine Cattaert Megrat [Inria, until Oct 2020]

Visiting Scientist

- Thomas Prieto Rumeau [Université nationale d'enseignement à distance - Espagne, until Jan 2020]

External Collaborators

- Dann Laneuville [DCNS Group, from Jul 2020]
- Benoite de Saporta [Univ de Montpellier, HDR]

2 Overall objectives

2.1 Presentation

The core component of our scientific agenda focuses on the development of statistical and probabilistic methods for the modeling and the optimization of complex systems. These systems require dynamic and stochastic mathematical representations with discrete and/or continuous variables. Their complexity poses genuine scientific challenges that can be addressed through complementary approaches and methodologies:

- *Modeling:* design and analysis of realistic and tractable models for such complex real-life systems taking into account various probabilistic phenomena;
- *Estimation:* developing theoretical and computational methods in order to estimate the parameters of the model and to evaluate the performance of the system;
- *Control:* developing theoretical and numerical control tools to optimize the performance.

These three approaches are strongly connected and the most important feature of the team is to consider these topics as a whole. This enables the team to deal with real industrial problems in several contexts such as biology, production planning, trajectory generation and tracking, performance and reliability.

3 Research program

3.1 Introduction

The scientific objectives of the team are to provide mathematical tools for modeling and optimization of complex systems. These systems require mathematical representations which are in essence dynamic, multi-model and stochastic. This increasing complexity poses genuine scientific challenges in the domain of modeling and optimization. More precisely, our research activities are focused on stochastic optimization and (parametric, semi-parametric, multidimensional) statistics which are complementary and interlinked topics. It is essential to develop simultaneously statistical methods for the estimation and control methods for the optimization of the models.

3.2 Main research topics

Stochastic modeling: Markov chain, Piecewise Deterministic Markov Processes (PDMP), Markov Decision Processes (MDP).

The mathematical representation of complex systems is a preliminary step to our final goal corresponding to the optimization of its performance. The team CQFD focuses on two complementary types of approaches. The first approach is based on mathematical representations built upon physical models where the dynamic of the real system is described by *stochastic processes.* The second one consists in studying the modeling issue in an abstract framework where the real system is considered as black-box. In this context, the outputs of the system are related to its inputs through a *statistical model.* Regarding stochastic processes, the team studies Piecewise Deterministic Markov Processes (PDMPs) and Markov Decision Processes (MDPs). These two classes of Markov processes form general families of controlled stochastic models suitable for the design of sequential decision-making problems. They appear in many fields such as biology, engineering, computer science, economics, operations research and provide powerful classes of processes for the modeling of complex systems. Our contribution to this topic consists in expressing real-life industrial problems into these mathematical frameworks. Regarding statistical methods, the team works on dimension reduction models. They provide a way to understand and visualize the structure of complex data sets. Furthermore, they are important tools in several different areas such as data analysis and machine learning, and appear in many applications such as biology, genetics, environment and recommendation systems. Our contribution to this topic consists in studying semiparametric modeling which combines the advantages of parametric and nonparametric models.

Estimation methods: estimation for PDMP; estimation in non- and semi-parametric regression modeling.
To the best of our knowledge, there does not exist any general theory for the problems of estimating parameters of PDMPs although there already exists a large number of tools for sub-classes of PDMPs such as point processes and marked point processes. To fill the gap between these specific models and the general class of PDMPs, new theoretical and mathematical developments will be on the agenda of the whole team. In the framework of non-parametric regression or quantile regression, we focus on kernel estimators or kernel local linear estimators for complete data or censored data. New strategies for estimating semi-parametric models via recursive estimation procedures have also received an increasing interest recently. The advantage of the recursive estimation approach is to take into account the successive arrivals of the information and to refine, step after step, the implemented estimation algorithms. These recursive methods do require restarting calculation of parameter estimation from scratch when new data are added to the base. The idea is to use only the previous estimations and the new data to refresh the estimation. The gain in time could be very interesting and there are many applications of such approaches.

Dimension reduction: dimension-reduction via SIR and related methods, dimension-reduction via multidimensional and classification methods.

Most of the dimension reduction approaches seek for lower dimensional subspaces minimizing the loss of some statistical information. This can be achieved in modeling framework or in exploratory data analysis context.
In modeling framework we focus our attention on semi-parametric models in order to conjugate the advantages of parametric and non-parametric modeling. On the one hand, the parametric part of the model allows a suitable interpretation for the user. On the other hand, the functional part of the model offers a lot of flexibility. In this project, we are especially interested in the semi-parametric regression model $Y = f(X'\theta) + \varepsilon$, the unknown parameter θ belongs to $\mathbb{R}^p$ for a single index model, or is such that $\theta = [\theta_1, \ldots, \theta_d]$ (where each θ_k belongs to $\mathbb{R}^p$ and $d \leq p$ for a multiple indices model), the noise ε is a random error with unknown distribution, and the link function f is an unknown real valued function. Another way to see this model is the following: the variables X and Y are independent given $X'\theta$. In our semi-parametric framework, the main objectives are to estimate the parametric part θ as well as the non-parametric part which can be the link function f, the conditional distribution function of Y given X or the conditional quantile q_α. In order to estimate the dimension reduction parameter θ we focus on the Sliced Inverse Regression (SIR) method which has been introduced by Li [39] and Duan and Li [37].
Methods of dimension reduction are also important tools in the field of data analysis, data mining and machine learning. They provide a way to understand and visualize the structure of complex data sets. Traditional methods among others are principal component analysis for quantitative variables or multiple component analysis for qualitative variables. New techniques have also been proposed to address these challenging tasks involving many irrelevant and redundant variables and often comparably few observation units. In this context, we focus on the problem of synthetic variables construction, whose goals include increasing the predictor performance and building more compact variables subsets. Clustering of variables is used for feature construction. The idea is to replace a group of ”similar” variables by a cluster centroid, which becomes a feature. The most popular algorithms include K-means and hierarchical clustering. For a review, see, e.g., the textbook of Duda [38].

Stochastic control: optimal stopping, impulse control, continuous control, linear programming.

The main objective is to develop *approximation techniques* to provide quasi-optimal feasible solutions and to derive *optimality results* for control problems related to MDPs and PDMPs:

- *Approximation techniques*. The analysis and the resolution of such decision models mainly rely on the maximum principle and/or the dynamic/linear programming techniques together with their various extensions such as the value iteration (VIA) and the policy iteration (PIA) algorithm. However, it is well known that these approaches are hardly applicable in practice and suffer from the so-called *curse of dimensionality*. Hence, solving numerically a PDMP or an MDP is a difficult and important challenge. Our goal is to obtain results which are both consistent from a theoretical point of view and computationally tractable and accurate from an application standpoint. It is important to emphasize that these research objectives were not planned in our initial 2009 program.

Our objective is to propose approximation techniques to efficiently compute the optimal value function and to get quasi-optimal controls for different classes of constrained and unconstrained MDPs with general state/action spaces, and possibly unbounded cost function. Our approach is based on combining the linear programming formulation of an MDP with probabilistic approximation techniques related to quantization techniques and the theory of empirical processes. An other aim is to apply our methods to specific industrial applications in collaboration with industrial partners such as Airbus Defence & Space, Naval Group and Thales.

Asymptotic approximations are also developed in the context of queueing networks, a class of models where the decision policy of the underlying MDP is in some sense fixed a priori, and our main goal is to study the transient or stationary behavior of the induced Markov process. Even though the decision policy is fixed, these models usually remain intractable to solve. Given this complexity, the team has developed analyses in some limiting regime of practical interest, i.e., queueing models in the large-network, heavy-traffic, fluid or mean-field limit. This approach is helpful to obtain a simpler mathematical description of the system under investigation, which is often given in terms of ordinary differential equations or convex optimization problems.

- *Optimality results.* Our aim is to investigate new important classes of optimal stochastic control problems including constraints and combining continuous and impulse actions for MDPs and PDMPs. In this framework, our objective is to obtain different types of optimality results. For example, we intend to provide conditions to guarantee the existence and uniqueness of the optimality equation for the problem under consideration and to ensure existence of an optimal (and ϵ-optimal) control strategy. We also plan to analyze the structural properties of the optimal strategies as well as to study the associated infinite dimensional linear programming problem. These results can be seen as a first step toward the development of numerical approximation techniques in the sense described above.

4 Application domains

4.1 Dependability and safety

Our abilities in probability and statistics apply naturally to industry, in particular in studies of dependability and safety. An illustrative example is the collaboration that started in September 2014 with with THALES Optronique. The goal of this project is the optimization of the maintenance of an onboard system equipped with a HUMS (Health Unit Monitoring Systems). The physical system under consideration is modeled by a piecewise deterministic Markov process. In the context of impulse control, we propose a dynamic maintenance policy, adapted to the state of the system and taking into account both random failures and those related to the degradation phenomenon.

The spectrum of applications of the topics that the team can address is large and can concern many other fields. Indeed non parametric and semi-parametric regression methods can be used in biometry, econometrics or engineering for instance. Gene selection from microarray data and text categorization are two typical application domains of dimension reduction among others. We had for instance the opportunity via the scientific program PRIMEQUAL to work on air quality data and to use dimension reduction techniques as principal component analysis (PCA) or positive matrix factorization (PMF) for pollution sources identification and quantization.

5 Highlights of the year

- 2020 was marked by the covid crisis and its impact on society and its activity. The world of research has also been greatly affected: faculty members have seen their teaching load increase significantly; PhD students and post-docs have often had to deal with a worsening of their working conditions, as well as reduced interactions with their supervisors and colleagues; most scientific collaborations have been greatly affected, with several international activities cancelled or postponed to dates still to be defined.

- During 2020, the group CQFD worked on the project of building a joint team with Naval Group. It constits on the first joint team with industry inside Inria.

6 New software and platforms

6.1 New software

6.1.1 vimplclust

Keyword: Clustering

Functional Description: vimpclust is an R package that implements methods related to sparse clustering and variable importance. The package currently allows to perform sparse k-means clustering with a group penalty, so that it automatically selects groups of numerical features. It also allows to perform sparse clustering and variable selection on mixed data (categorical and numerical features), by preprocessing each categorical feature as a group of numerical features. Several methods for visualizing and exploring the results are also provided.

URL: `https://CRAN.R-project.org/package=vimpclust`

Contacts: Marie Chavent, Alex Mourer

6.1.2 dd-sPLS

Name: Data-Driven Sparse PLS

Keywords: Marker selection, Classification, Regression, Missing data, Multi-Block, High Dimensional Data, PLS, SVD, Sparsity

Scientific Description: Allows to build Multi-Data-Driven Sparse PLS models. Multi-blocks with high-dimensional settings are particularly sensible to this. Whatsmore it deals with missing samples (entire lines missing per block) thanks to the Koh-Lanta algorithm. SVD decompositions permit to offer a fast and controlled method.

Functional Description: ddsPLS is an R package that proposes a sparse PLS formulation for mono and multi-block data sets with or without missing samples. ddsPLS stands for data-driven sparse partial least square, allowing variable selection in the covariate and the response blocks in the context of numerous variables and a low number of individuals.

URL: `https://github.com/hlorenzo/ddsPLS`

Contacts: Hadrien Lorenzo, Jérôme Saracco, Rodolphe Thiebaut

6.1.3 outlierSIR

Keyword: Regression

Functional Description: outlierSIR is an R package R that proposes thee computational methods to detect outliers in a single-index regression model estimated with SIR (sliced inverse regression) approach and kernel regression. Three outlier detection methods are available: a naive method called MONO, a method that relies on training sample and test sample replications for evaluating the "stability" of the estimated model, named TTR (for for Training Test Replications), and the BOOT method based on Bootstap replications and in-bag errors.

URL: `https://github.com/hlorenzo/outlierSIR/`

Contacts: Hadrien Lorenzo, Jérôme Saracco

7 New results

7.1 Power-of-d-Choices with Memory: Fluid Limit and Optimality

In multi-server distributed queueing systems, the access of stochastically arriving jobs to resources is often regulated by a dispatcher, also known as load balancer. A fundamental problem consists in designing a load balancing algorithm that minimizes the delays experienced by jobs. During the last two decades, the power-of-d-choice algorithm, based on the idea of dispatching each job to the least loaded server out of servers randomly sampled at the arrival of the job itself, has emerged as a break through in the foundations of this area due to its versatility and appealing asymptotic properties. In this paper [6], we consider the power-of-d-choice algorithm with the addition of a local memory that keeps track of the latest observations collected over time on the sampled servers. Then, each job is sent to a server with the lowest observation. We show that this algorithm is asymptotically optimal in the sense that the load balancer can always assign each job to an idle server in the large-system limit. Our results quantify and highlight the importance of using memory as a mean to enhance performance in randomized load balancing.

Authors: J. Anselmi (CQFD); F. Dufour (CQFD).

7.2 A Convex Programming Approach for Discrete-Time Markov Decision Processes under the Expected Total Reward Criterion

In this work [14], we study discrete-time Markov decision processes (MDPs) under constraints with Borel state and action spaces and where all the performance functions have the same form of the expected total reward (ETR) criterion over the infinite time horizon. One of our objective is to propose a convex programming formulation for this type of MDP. It will be shown that the values of the constrained control problem and the associated convex program coincide. Moreover, if there exists an optimal solution to the convex program then there exists a stationary randomized policy which is optimal for the MDP. It will be also shown that in the framework of constrained control problems, the supremum of the ETRs over the set of randomized policies is equal to the supremum of the ETRs over the set of stationary randomized policies. We consider standard hypotheses such as the so-called continuity-compactness conditions and a Slater-type condition. Our assumptions are quite weak to deal with cases that have not yet been addressed in the literature. Examples are presented to illustrate our results.

Authors: F. Dufour (CQFD); A. Genadot (CQFD).

7.3 On the expected total cost with unbounded returns for Markov decision processes

In this work [15], we consider a discrete-time Markov decision process with Borel state and action spaces. The performance criterion is to maximize a total expected utility determined by unbounded return function. It is shown the existence of optimal strategies under general conditions allowing the reward function to be unbounded both from above and below and the action sets available at each step to the decision maker to be not necessarily compact. To deal with unbounded reward functions, a new characterization for the weak convergence of probability measures is derived. Our results are illustrated by examples.

Authors: F. Dufour (CQFD); A. Genadot (CQFD).

7.4 A backward Itô–Ventzell formula with an application to stochastic interpolation

This note [13] and its extended version (see reference [7] of this note) present a novel backward Itô–Ventzell formula and an extension of the Aleeksev–Gröbner interpolating formula to stochastic flows. We also present some natural spectral conditions that yield direct and simple proofs of time uniform estimates of the difference between the two stochastic flows when their drift and diffusion functions are not the same.

Authors: P. Del Moral (CQFD); S. Singh.

7.5 A duality formula and a particle Gibbs sampler for continuous time Feynman-Kac measures on path spaces

Continuous time Feynman-Kac measures on path spaces are central in applied probability, partial differential equation theory, as well as in quantum physics. This article [7] presents a new duality formula between normalized Feynman-Kac distribution and their mean field particle interpretations. Among others, this formula allows us to design a reversible particle Gibbs-Glauber sampler for continuous time Feynman-Kac integration on path spaces. We also provide new Dyson-Phillips semigroup expansions, as well as novel uniform propagation of chaos estimates for continuous time genealogical tree based particle models with respect to the time horizon and the size of the systems. Our approach is self contained and it is based on a novel stochastic perturbation analysis and backward semigroup techniques. These techniques allow to obtain sharp quantitative estimates of the convergence rate to equilibrium of particle Gibbs-Glauber samplers. To the best of our knowledge these results are the first of this kind for continuous time Feynman-Kac measures.

Authors: M. Arnaudon; P. Del Moral (CQFD).

7.6 A second order analysis of McKean–Vlasov semigroups

We propose in [8] a second order differential calculus to analyze the regularity and the stability properties of the distribution semigroup associated with McKean–Vlasov diffusions. This methodology provides second order Taylor type expansions with remainder for both the evolution semigroup as well as the stochastic flow associated with this class of nonlinear diffusions. Bismut–Elworthy–Li formulae for the gradient and the Hessian of the integro-differential operators associated with these expansions are also presented.

The article also provides explicit Dyson–Phillips expansions and a refined analysis of the norm of these integro-differential operators. Under some natural and easily verifiable regularity conditions we derive a series of exponential decays inequalities with respect to the time horizon. We illustrate the impact of these results with a second order extension of the Alekseev–Gröbner lemma to nonlinear measure valued semigroups and interacting diffusion flows. This second order perturbation analysis provides direct proofs of several uniform propagation of chaos properties w.r.t. the time parameter, including bias, fluctuation error estimate as well as exponential concentration inequalities.

Authors: M. Arnaudon; P. Del Moral (CQFD).

7.7 A perturbation analysis of stochastic matrix Riccati diffusions

Matrix differential Riccati equations are central in filtering and optimal control theory. The purpose of this article is to develop a perturbation theory for a class of stochastic matrix Riccati diffusions. Diffusions of this type arise, for example, in the analysis of ensemble Kalman–Bucy filters since they describe the flow of certain sample covariance estimates. In this context, the random perturbations come from the fluctuations of a mean field particle interpretation of a class of nonlinear diffusions equipped with an interacting sample covariance matrix functional. The main purpose of this article [10] is to derive non-asymptotic Taylor-type expansions of stochastic matrix Riccati flows with respect to some perturbation parameter. These expansions rely on an original combination of stochastic differential analysis and nonlinear semigroup techniques on matrix spaces. The results here quantify the fluctuation of the stochastic flow around the limiting deterministic Riccati equation, at any order. The convergence of the interacting sample covariance matrices to the deterministic Riccati flow is proven as the number of particles tends to infinity. Also presented are refined moment estimates and sharp bias and variance estimates. These expansions are also used to deduce a functional central limit theorem at the level of the diffusion process in matrix spaces.

Authors: A. Bishop; P. Del Moral (CQFD); and A. Niclas.

7.8 Applicability and Interpretability of Ward's Hierarchical Agglomerative Clustering With or Without Contiguity Constraints

Hierarchical agglomerative clustering (HAC) with Ward's linkage has been widely used since its introduction by Ward (Journal of the American Statistical Association, 58(301), 236–244, 1963). This article reviews extensions of HAC to various input data and contiguity-constrained HAC, and provides applicability conditions. In addition, different versions of the graphical representation of the results as a dendrogram are also presented and their properties are clarified. We clarify and complete the results already available in an heterogeneous literature using a uniform background. In particular, this study reveals an important distinction between a consistency property of the dendrogram and the absence of crossover within it. Finally, a simulation study shows that the constrained version of HAC can sometimes provide more relevant results than its unconstrained version despite the fact that the constraint leads to optimize the objective criterion on a reduced set of solutions at each step. Overall, this article provides comprehensive recommendations, both for the use of HAC and constrained HAC depending on the input data, and for the representation of the results [18].

Authors: Randriamihamison (CQFD); Nathalie Vialaneix; P. Neuvial.

7.9 La modélisation de l'indemnisation du préjudice corporel : Un exemple de «justice quantitative» au service de l'équité

Les pratiques judiciaires relatives à un contentieux permettent de révéler les difficultés auxquelles sont confrontés les magistrats. Concernant la réparation d'un dommage corporel, ces difficultés sont nombreuses puisque les magistrats doivent composer avec le sacro-saint principe de la réparation intégrale d'une part, et l'absence de consensus sur la méthodologie d'autre part. Les statistiques permettent de mettre en évidence ces hésitations, et plus particulièrement une certaine hétérogénéité des indemnisations d'un préjudice corporel. De toute évidence, la nature de ce contentieux, à savoir le dommage fait au corps, revêt une dimension particulièrement humaine et délicate. La difficulté majeure étant l'indemnisation des préjudices extrapatrimoniaux. La nécessité d'une équité des victimes devant la réparation d'un préjudice corporel est impérieuse. Les pratiques judiciaires révèlent des différences dans l'indemnisation qui semblent être corrélées entre autres à la cour d'appel statuant sur la demande de réparation du dommage corporel. Comme nous le montrons dans [30], la modélisation mathématique du processus de décision permettrait de dépasser un référentiel condamné à demeurer obsolète puisqu'il se réfère à des statistiques qui ne sont qu'une photographie figée de la jurisprudence passée.

Authors: Anaïs Gayte-Papon de Lameigné, Pierrick Legrand (CQFD), Jacques Levy-Vehel

7.10 EEG Feature Extraction Using Genetic Programming for the Classification of Mental States

The design of efficient electroencephalogram (EEG) classification systems for the detection of mental states is still an open problem. Such systems can be used to provide assistance to humans in tasks where a certain level of alertness is required, like in surgery or in the operation of heavy machines, among others. In this work [21], we extend a previous study where a classification system is proposed using a Common Spatial Pattern (CSP) and Linear Discriminant Analysis (LDA) for the classification of two mental states, namely a relaxed and a normal state. Here, we propose an enhanced feature extraction algorithm (Augmented Feature Extraction with Genetic Programming, or+FEGP) that improves upon previous results by employing a Genetic-Programming-based methodology on top of the CSP. The proposed algorithm searches for non-linear transformations that build new features and simplify the classification task. Although the proposed algorithm can be coupled with any classifier, LDA achieves 78.8 percent of accuracy, the best predictive accuracy among tested classifiers,significantly improving upon previously published results on the same real-world dataset.

Authors: Emigdio Z-Flores, Leonardo Trujillo, Pierrick Legrand (CQFD), Frédérique Faïta-Aïnseba

7.11 Alternative Ways to Compare the Detendred Fluctuation Analysis and its Variants. Application to Visual Tunneling Detection

The detrended fluctuation analysis (DFA) and its variants such as the detrended moving average (DMA) are widely used to estimate the Hurst exponent. These methods are very popular as they do not require advanced skills in the field of signal processing and statistics while providing accurate results. As a consequence, a great deal of interest has been paid to compare them and to better understand their behaviors from a mathematical point of view. In this work [9], our contribution is threefold. Firstly, we propose another variant avoiding the discontinuities between consecutive local trends of the DFA by a priori constraining them to be continuous. Secondly, we show that, in all these approaches, the square of the fluctuation function can be presented in a similar matrix form. When the process is wide-sense stationary (w.s.s.), the latter can be seen as the power of the output of a linear filtering whose frequency response depends on the given method. In the general case, an interpretation of the square of the fluctuation function is also given by expressing it as the convolution between the 2D-Fourier transform of two matrices, one whose elements correspond to the instantaneous correlation function of the signal and the other which depends on the detrending method. To end up, an illustration is provided in the field of avionics for the detection of the visual tunneling, a deleterious cognitive state.

Authors: Bastien Berthelot, Eric Grivel, Pierrick Legrand (CQFD), Jean-Marc André, Patrick Mazoyer

7.12 SOAP: Semantic Outliers Automatic Preprocessing

Genetic Programming (GP) is an evolutionary algorithm for the automatic generation of symbolic models expressed as syntax trees. GP has been successfully applied in many domain, but most research in this area has not considered the presence of outliers in the training set. Outliers make supervised learning problems difficult, and sometimes impossible, to solve. For instance, robust regression methods cannot handle more than 50 percent of outlier contamination, referred to as their breakdown point. This work [20] studies problems where outlier contamination is high, reaching up to 90 percent contamination levels, extreme cases that can appear in some domains. This work shows, for the first time, that a random population of GP individuals can detect outliers in the output variable. From this property, a new filtering algorithm is proposed called Semantic Outlier Automatic Preprocessing (SOAP), which can be used with any learning algorithm to differentiate between inliers and outliers. Since the method uses a GP population, the algorithm can be carried out for free in a GP symbolic regression system. The approach is the only method that can perform such an automatic cleaning of a dataset without incurring an exponential cost as the percentage of outliers in the dataset increases.

Authors: Leonardo Trujillo, Uriel Lopez, Pierrick Legrand (CQFD)

7.13 A Variable Chirp Rate Stepped Frequency Linear Frequency Modulation Waveform Designed to Approximate Wideband Non-Linear Radar Waveforms

The non-linear frequency modulation (NLFM) waveform is one of the existing waveforms that can be used in high range resolution radar applications. However, a high sampling frequency and consequently an expensive ADC are required. To overcome this drawback while taking advantage of the features of the NLFM waveform, we suggest approximating the wideband NLFM waveform by a piecewise linear waveform and using it in a stepped frequency (SF) framework. Thus, a variable chirp rate SF-LFM waveform is proposed where SF is combined with a train of LFM pulses having different chirp rates, durations, and bandwidths. In this work [19], these parameters are derived from a tangent-based NLFM waveform. At the receiver, a generalized version of the time domain (TD) algorithm is proposed to process the received echoes. Our purpose is to obtain the high range resolution profile (HRRP) whose properties are of the same magnitude orders as those obtained using a tangent-based NLFM waveform. These properties are the peak sidelobe ratio, the integrated sidelobe, and the range resolution. Toward this goal, a multi-objective optimization issue is addressed to deduce the parameters of the proposed waveform by using two types of approaches based on evolutionary algorithms. Their relevance is compared. Our analysis and simulations show that the proposed approaches attain the targeted performance goals with a smaller sampling frequency at the receiver.

Authors: Mahdi Saleh, Samir-Mohamad Omar, Eric Grivel, Pierrick Legrand (CQFD)

7.14 Sea Target Classification Based On An A Priori Motion Model

Target classification can be of real interest for sea surveillance in both civil and military contexts. To address this issue, we present two approaches based on the Singer model [24]. The latter has the advantage of covering a wide range of motions depending on the values of its parameters. Given noisy observations, the first method aims at estimating the motion model parameters by taking advantage of the properties of the correlation function of the estimated acceleration. It is based on a genetic algorithm. The second approach is on-line and consists in deriving a joint tracking and classification (JTC) method. Based on various simulations, we study their respective relevance in different operational settings. The proposed JTC corresponds to the best compromise in terms of performance and number of samples required.

Authors: Jimmy Bondu, Eric Grivel, Audrey Giremus, Pierrick Legrand (CQFD), Vincent Corretja, et al.

7.15 Regularized Dfa To Study The Gaze Position Of An Airline Pilot

To estimate the Hurst exponent of a mono-fractal process, the detrended fluctuation analysis (DFA) is based on the estimation of the trend of the integrated process. The latter is subtracted from the integrated process. The power of the residual is then computed and corresponds to the square of the fluctuation function. Its logarithm is proportional to the Hurst exponent. In the last few years, a few variants of this method have been proposed and differ in the way of estimating the trend. Our contribution in this work [23] is threefold. First, we introduce a new variant of the DFA, based on a regularized least-square criterion to estimate the trend. Then, the influence of the regularization parameter on the fluctuation function is analyzed in two cases: when the process is wide sense stationary and when it is not. Finally, an application is presented in the field of aeronautics to characterize an attentional impairment: the visual tunneling.

Authors: Bastien Berthelot, Eric Grivel, Pierrick Legrand (CQFD), Jean-Marc André, Patrick Mazoyer

7.16 Unlabeled Multi-Target Regression with Genetic Programming

Machine Learning (ML) has now become an important and ubiquitous tool in science and engineering, with successful applications in many real-world domains. However, there are still areas in need of improvement, and problems that are still considered difficult with off-the-shelf methods. One such problem is Multi Target Regression (MTR), where the target variable is a multidimensional tuple instead of a scalar value. In this work [25], we propose a more difficult variant of this problem which we call Unlabeled MTR (uMTR), where the structure of the target space is not given as part of the training data. This version of the problem lies at the intersection of MTR and clustering, an unexplored problem type. Moreover, this work proposes a solution method for uMTR, a hybrid algorithm based on Genetic Programming and RANdom SAmple Consensus (RANSAC). Using a set of benchmark problems, we are able to show that this approach can effectively solve the uMTR problem.

Authors: Uriel Lopez, Leonardo Trujillo, Sara Silva, Leonardo Vanneschi, Pierrick Legrand (CQFD).

7.17 Reconstructing climatic modes of variability from proxy records using ClimIndRec version 1.0

Modes of climate variability strongly impact our climate and thus human society. Nevertheless, the statistical properties of these modes remain poorly known due to the short time frame of instrumental measurements. Reconstructing these modes further back in time using statistical learning methods applied to proxy records is useful for improving our understanding of their behaviour. For doing so, several statistical methods exist, among which principal component regression is one of the most widely used in paleoclimatology. In this work [17], we provide the software ClimIndRec to the climate community; it is based on four regression methods (principal component regression, PCR; partial least squares, PLS; elastic net, Enet; random forest, RF) and cross-validation (CV) algorithms, and enables the systematic reconstruction of a given climate index. A prerequisite is that there are proxy records in the database that overlap in time with its observed variations. The relative efficiency of the methods can vary, according to the statistical properties of the mode and the proxy records used. Here, we assess the sensitivity to the reconstruction technique. ClimIndRec is modular as it allows different inputs like the proxy database or the regression method. As an example, it is here applied to the reconstruction of the North Atlantic Oscillation by using the PAGES 2k database. In order to identify the most reliable reconstruction among those given by the different methods, we use the modularity of ClimIndRec to investigate the sensitivity of the methodological setup to other properties such as the number and the nature of the proxy records used as predictors or the targeted reconstruction period. We obtain the best reconstruction of the North Atlantic Oscillation (NAO) using the random forest approach. It shows significant correlation with former reconstructions, but exhibits higher validation scores.

Authors: Simon Michel, Didier Swingedouw, Marie Chavent (CQFD), Pablo Ortega, Juliette Mignot, et al.

7.18 Various Statistical Approaches to Assess and Predict Carcass and Meat Quality Traits

The beef industry is organized around different stakeholders, each with their own expectations, sometimes antagonistic. This work [16] first outlines these differing perspectives. Then, various optimization models that might integrate all these expectations are described. The final goal is to define practices that could increase value for animal production, carcasses and meat whilst simultaneously meeting the main expectations of the beef industry. Different models previously developed worldwide are proposed here. Two new computational methodologies that allow the simultaneous selection of the best regression models and the most interesting covariates to predict carcass and/or meat quality are developed. Then, a method of variable clustering is explained that is accurate in evaluating the interrelationships between different parameters of interest. Finally, some principles for the management of quality trade-offs are presented and the Meat Standards Australia model is discussed. The “Pareto front” is an interesting approach to deal jointly with the different sets of expectations and to propose a method that could optimize all expectations together.

Authors: Marie-Pierre Ellies-Oury, Jean-François Hocquette, Sghaier Chriki, Alexandre Conanec, Linda Farmer, Marie Chavent (CQFD), Jérôme Saracco (CQFD)

7.19 Multiple-output quantile regression through optimal quantization

A new nonparametric quantile regression method based on the concept of optimal quantization was developed recently and was showed to provide estimators that often dominate their classical, kernel-type, competitors. In the present work [11], we extend this method to multiple-output regression problems. We show how quantization allows approximating population multiple-output regression quantiles based on halfspace depth. We prove that this approximation becomes arbitrarily accurate as the size of the quantization grid goes to infinity. We also derive a weak consistency result for a sample version of the proposed regression quantiles. Through simulations, we compare the performances of our estimators with (local constant and local bilinear) kernel competitors. The results reveal that the proposed quantization-based estimators, which are local constant in nature, outperform their kernel counterparts and even often dominate their local bilinear kernel competitors. The various approaches are also compared on artificial and real data.

Authors: Isabelle Charlier, Davy Paindaveine, Jérôme Saracco(CQFD)

7.20 BIG-SIR: a Sliced Inverse Regression approach for massive data

In a massive data setting, we focus on a semiparametric regression model involving a real dependent variable Y and a p-dimensional covariate X (with $p \geq 1$). This model includes a dimension reduction of X via an index $X'\beta$. The Effective Dimension Reduction (EDR) direction β cannot be directly estimated by the Sliced Inverse Regression (SIR) method due to the large volume of the data. To deal with the main challenges of analyzing massive data sets which are the storage and computational efficiency, we propose in this work [28] a new SIR estimator of the EDR direction by following the "divide and conquer" strategy. The data is divided into subsets. EDR directions are estimated in each subset which is a small data set. The recombination step is based on the optimization of a criterion which assesses the proximity between the EDR directions of each subset. Computations are run in parallel with no communication among them. The consistency of our estimator is established and its asymptotic distribution is given. Extensions to multiple indices models, q-dimensional response variable and/or SIRα- based methods are also discussed. A simulation study using our edr Graphical Tools R package shows that our approach enables us to reduce the computation time and conquer the memory constraint problem posed by massive data sets. A combination of foreach and big memory R packages are exploited to offer efficiency of execution in both speed and memory. Results are visualized using the bin-summarises mooth approach through the bigvis R package. Finally, we illustrate our proposed approach on a massive airline data set.

Authors: Benoit Liquet, Jérôme Saracco(CQFD)

7.21 A computational methodology for multi-objective optimization in precision rearing

In precision rearing, optimization problems are multi-objective and stochastic because the requirements of decision-makers are multiple and because objective functions cannot be modeled in an analytical form due to the inherent complexity of biological systems. Our method [27] consists in use a nonparametrically estimated quantile regression, associated with a α risk level decided by the decision maker, to deal with the model uncertainty. Then, the NSGAII genetic algorithm allows us to find the Pareto Front, associated with a α risk level, which carries the set of possible trade-offs within which the decision-maker can choose. The good numerical behavior of the proposed approach is illustrated on simulated data. Keywords: Optimization, Multi-objectives, Uncertainty, Precision rearing, Conditional quantiles.

Authors: Alexandre Conanec, Marie Chavent (CQFD), Marie Pierre Ellies-Oury, Jérôme Saracco (CQFD)

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

Naval Group Participants: Huilong Zhang, Pierre Del Moral, Dann Laneuville, Alexandre Genadot, François Dufour.

The increasing complexity of warfare submarine missions has led Naval Group to study new tactical help functions for underwater combat management systems. In this context, the objective is to find optimal trajectories according to the current mission type by taking into account sensors, environment and surrounding targets. This problem has been modeled as a discrete-time Markov decision process with finite horizon. Different kind of scenarios have been considered and studied.

Thales Optronique Participants: Benoîte de Saporta, François Dufour, Tiffany Cerchi.

Maintenance, optimization, fleet of industrial equipements. The topic of this collaboration with Université de Montpellier and Thales Optronique is the application of Markov decision processes to the maintenance optimization of a fleet of industrial equipments.

Thales AVS Participants: Bastien Berthelot, Pierrick Legrand.
The collaboration is centered around some contributions to the estimation of the Hurst coefficient and his application on biosignals in the domain of crew monitoring.

Case Law Analytics Participant: Pierrick Legrand.
Pierrick Legrand is a consultant for the startup Case Law Analytics. The object of the consulting is confidential.

Sartorius Participant: Hadrien Lorenzo Jérôme Saracco.

The team is currently initiating a scientific collaboration with the Advanced Data Analytics Group of Sartorius Corporate Research which is an international pharmaceutical and laboratory equipment supplier, covering the segments of Bioprocess Solutions and Lab Products and Services. The current work concerns the development of a PLS (Partial Least Squares) inspired method in the context of multiblock of covariates (corresponding to different technologies and/or different sampling, statistical natures...) and high dimensional datasets (with the sample size n much smaller than the number of variables in the different blocks). The proposed method, called ddsPLS for data-driven sparse PLS, allows variable selection in the X and in the Y parts thanks to interpretable parameters associate with the soft-thresolding of the empirical correlation matrices (between the X's blocks and the Y block) decomposed in SVD (Singular Values Decomposition) ways. In addition a methodology to handle specific missing values (i.e. missing samples in some covariate blocks) is also under investigation.

Safran Aircraft Engines Participant: Marie Chavent, Jérôme Lacaille, Alex Mourer, Madalina Olteanou

The collaboration is centered around an applied mathematics thesis defining a formalism and a methodology for processing and interpretation by the importance of variables (from measurements and calculated indicators) in the case of unsupervised problems. This methodology is accompanied by code programming and a demonstration on an example data set from Safran Aircraft Engines.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Inria international partners

Declared Inria international partners: P. Legrand is working with L. Trujillo (ITT, Tijuana).

Informal international partners: J. Saracco is working with Prof. Davy Paindaveine (ULB, Bruxelles, Belgium).

F. Dufour is working with O.L.V. Costa (University of Sao Paulo, Brasil).

F. Dufour is working with A. Piunovskiy (University of Liverpool, UK).

F. Dufour is working with T. Prieto-Rumeau (UNED, Spain).

9.2 National initiatives

QuAMProcs of the program *Project Blanc* of the ANR The mathematical analysis of metastable processes started 75 years ago with the seminal works of Kramers on Fokker-Planck equation. Although the original motivation of Kramers was to « elucidate some points in the theory of the velocity of chemical reactions », it turns out that Kramers' law is observed to hold in many scientific fields: molecular biology (molecular dynamics), economics (modelization of financial bubbles), climate modeling, etc. Moreover, several widely used efficient numerical methods are justified by the mathematical description of this phenomenon.

Recently, the theory has witnessed some spectacular progress thanks to the insight of new tools coming from Spectral and Partial Differential Equations theory.

Semiclassical methods together with spectral analysis of Witten Laplacian gave very precise results on reversible processes. From a theoretical point of view, the semiclassical approach allowed to prove a complete asymptotic expansion of the small eigen values of Witten Laplacian in various situations (global problems, boundary problems, degenerate diffusions, etc.). The interest in the analysis of boundary problems was rejuvenated by recent works establishing links between the Dirichlet problem on a bounded domain and the analysis of exit event of the domain. These results open numerous perspectives of applications. Recent progress also occurred on the analysis of irreversible processes (e.g. on overdamped Langevin equation in irreversible context or full (inertial) Langevin equation).

The above progresses pave the way for several research tracks motivating our project: overdamped Langevin equations in degenerate situations, general boundary problems in reversible and irreversible case, non-local problems, etc.

Chaire Stress Test of the Ecole Polytechnique The Chaire "Stress Testing" is a specific research program between Ecole Polytechnique, BNP Paribas, Fondation de l'Ecole Polytechnique, and is hosted

at Polytechnique by the Center of Applied Mathematics. This research project is part of an in-depth reflection on the increasingly sophisticated issues surrounding stress tests (under the impulse of the upcoming European Banking regulation). Simulation of extreme adverse scenarios is an important topic to better understand which critical configurations can lead to financial and systemic crises. These scenarios may depend on complex phenomena, for which we partially lack information, making the modeling incomplete and uncertain. Last, the data are multivariate and reflect the dependency between driving variables. From the above observations, different lines of research are considered:
1. the generation of stress test and meta-modeling scenarios using machine learning;
2. the quantification of uncertainties in risk metrics;
3. modeling and estimation of multidimensional dependencies.

Mission pour les initiatives transverses et interdisciplinaires, Défi Modélisation du Vivant, projet MISGIVING The aim of MISGIVING (MathematIcal Secrets penGuins dIVING) is to use mathematical models to understand the complexity of the multiscale decision process conditioning not only the optimal duration of a dive but also the diving behaviour of a penguin inside a bout. A bout is a sequence of succesive dives where the penguin is chasing prey. The interplay between the chasing period (dives) and the resting period due to the physiological cost of a dive (the time spent at the surface) requires some kind of optimization.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair F. Dufour was the chair the selection committee for the 2021 SIAG/CST Best SICON Paper Prize. The deliberations of the committee took place in 2020.

Member of the organizing committees Marie Chavent is member of the Conference Programm Committee of the "8èmes Rencontres R" organised at AgroParisTech.

10.1.2 Scientific events: selection

Reviewer Pierrick Legrand is reviewer for the conferences GECCO, EUSIPCO and EA.

10.1.3 Journal

Member of the editorial boards J. Saracco is a member of the Editorial Board of Astrostatistics (specialty section of Frontiers in Astronomy and Space Sciences) since 2019.

Marie Chavent is a member of the editorial committee of the Pratique R collection at EDP Sciences

P. Del Moral is an associate editor for the journal Stochastic Analysis and Applications since 2001.

P. Del Moral is an associate editor for the journal Revista de Matematica: Teoria y aplicaciones since 2009.

P. Del Moral is an associate editor for the journal Annals of Applied Probability since 2019.

F. Dufour is corresponding editor of the SIAM Journal of Control and Optimization since 2018.

F. Dufour is associate editor of the journal Applied Mathematics and Optimization (AMO) since 2018.

F. Dufour is associate editor of the journal Stochastics: An International Journal of Probability and Stochastic Processes since 2018.

F. Dufour is the representative of the SIAM activity group in control and system theory for the journal SIAM News since 2014.

Pierrick Legrand is the main editor for the Springer LNCS volumes "Articicial Evolution" since 2009.

Reviewer - reviewing activities J. Saracco is a regular reviewer for many journal of statistics : The Annals of Statistics, Journal of Multivariate Analysis, Statistica Sinica, Biometrika, Communications in Statistics – Theory and Methods, Computational Statistics and Data Analysis, Journal of Statistical Planning and Inference, Computational Statistics, Journal of Machine Learning Research,...

Marie Chavent has been reviewer for PLOS one, Electronic Journal of Applied Statistical Analysis, Advances in Data Analysis and Classification

10.1.4 Leadership within the scientific community

Pierrick Legrand is the president of the association EA. https://sites.google.com/view/artificial-evolution/home

10.1.5 Scientific expertise

J. Saracco is an elected member of the CNU 26 (National Council of Universities in Applied Mathematics), since 2019.

10.1.6 Research administration

J. Saracco is the leader of the team OptimAl of Institut de Mathématiques de Bordeaux (IMB, UMR CNRS 5251) from 2019.

Marie Chavent is a member appointed to the Council of the Department of Engineering and Digital Sciences (SIN) of the University of Bordeaux.

10.2 Teaching - Supervision - Juries

Teaching

- J. Saracco is the head of the engineering department of ENSC, Graduate School of Cognitics (applied cognitive science and technology) which is a Bordeaux INP engineering school.
- Marie Chavent is in charge of the first year of the MIASHS degree program at the University of Bordeaux.
- Alexandre Genadot is in charge of the first year of the MIASHS degree program at the University of Bordeaux.
- Pierrick Legrand is in charge of the mathematics program for the MIASHS degree at the University of Bordeaux.
- Licence : P. Legrand, Algèbre, 129h, L1, Université de Bordeaux, France.
- Licence : P. Legrand, Espaces Euclidiens, 46,5h, L2, Université de Bordeaux, France.
- Licence : P. Legrand, Informatique pour les mathématiques, 30h, L2, Université de Bordeaux, France.
- DU : P. Legrand, Evolution Artificielle, Big data, 8h, DU, Bordeaux INP, France.
- Licence : A. Genadot, Bases en Probabilités, 18h, L1, Université de Bordeaux, France.
- Licence : A. Genadot, Projet Professionnel de l'étudiant, 8h, L1, Université de Bordeaux, France.
- Licence : A. Genadot, Probabilité, 30h, L2, Université de Bordeaux, France.
- Licence : A. Genadot, Techniques d'Enquêtes, 10h, L2, Université de Bordeaux, France.
- Licence : A. Genadot, Modélisation Statistiques, 16.5h, L3, Université de Bordeaux, France.
- Licence : A. Genadot, Préparation Stage, 15h, L3, Université de Bordeaux, France.

- Licence : A. Genadot, TER, 5h, L3, Université de Bordeaux, France.
- Licence : A. Genadot, Processus, 16.5h, L3, Université de Bordeaux, France.
- Licence : A. Genadot, Statistiques, 20h, L3, Bordeaux INP, France.
- Master : A. Genadot, Savoirs Mathématiques, 81h, M1, Université de Bordeaux et ESPE, France.
- Master : A. Genadot, Martingales, 29h, M1, Université de Bordeaux, France.
- Licence : F. Dufour, Probabilités et statistiques, 70h, first year of école ENSEIRB-MATMECA, Institut Polytechnique de Bordeaux, France.
- Master : F. Dufour, Approche probabiliste et methode de Monte Carlo, 24h, third year of école ENSEIRB-MATMECA, Institut Polytechnique de Bordeaux, France.
- Licence : J. Saracco, Probabilités et Statistique, 27h, first year of Graduate Schools of Engineering ENSC-Bordeaux INP, Institut Polytechnique de Bordeaux, France.
- Licence : J. Saracco, Statistique inférentielle et Analyse des données, 45h, first year of Graduate Schools of Engineering ENSC-Bordeaux INP, Institut Polytechnique de Bordeaux, France.
- Licence : J. Saracco, Statistique pour l'ingénieur, 16h, first year of Graduate Schools of Engineering ENSPIMA-Bordeaux INP, Institut Polytechnique de Bordeaux, France.
- Master : J. Saracco, Modélisation statistique, 81h, second year of Graduate Schools of Engineering ENSC-Bordeaux INP, Institut Polytechnique de Bordeaux, France.
- DU : J. Saracco, Statistique et Big data, 45h, DU BDSI (Big data et statistique pour l'ingenieur), Bordeaux INP, France.
- Licence : M. Chavent, Statistique Inférentielle, 18h, L2, Université de Bordeaux, France
- Licence : M. Chavent, Techniques d'Enquêtes, 10h, L2, Université de Bordeaux, France
- Master : M. Chavent, DataMining, 43h, M2, Université de Bordeaux
- Master : M. Chavent, Machine Learning, 58h, Université de Bordeaux,
- DU: M. Chavent, Apprentissage, 12h, DU BDSI, Bordeaux INP, France

Supervision

- PhD: Hadrien Lorenzo, "Supervised analysis of high dimensional multi block data", supervised by Jérôme Saracco (CQFD) and Rodolphe Thebaut (Inserm), thesis defense: 27/11/19 in Bordeaux.
- PhD in progress: Alex Mourer, "Variables importance in clustering", CIFRE Safran Aircraft Engines, supervised by Jérôme Lacaille (Safran), Madalina Olteanou (SAMM, Paris1), Alex Mourer (doctorant), Marie Chavent (CQFD).
- PhD in progress: de Nathanaël Randriamihamison, "Contiguity Constrained Hierarchical Agglomerative Clustering for Hi-C data analysis", supervised by Nathalie Vialaneix (MIAT, INRA Toulouse), Pierre Neuvial (IMT, CNRS), Marie Chavent (CQFD) .
- PhD in progress: Alexandre Conanec, "Modulation et optimisation statistique de données multi-tableaux : modélisation des facteurs de variations dans la gestion des compromis entre différents jeux de données", supervised by Marie Chavent(CQFD), Jérôme Saracco (CQFD), Marie-Pierre Ellies (INRA).
- PhD defended in october 2020: Loic Labache, "Création d'un atlas cérébral évolutif de régions fonctionnelles définies à partir d'une cohorte de 297 sujets ayant effectués 20 tâches cognitives en IRMf", supervised by Jérôme Saracco (CQFD), Marc Joliot (CEA).

- PhD in progress: Tiffany Cherchi, "Automated optimal fleet management policy for airborne equipment", Montpellier University, since 2017, supervised by B. De Saporta and F. Dufour.
- PhD in progress: Bastien Berthelot, "Contributions à l'estimation du coefficient de Hurst et son usage sur des biosignaux dans le domaine du crew monitoring", CIFRE THALES, supervised by P. Legrand. PhD defense: March 30, 2021.
- PhD in progress: Camille Palmier, "Nouvelles approches de fusion multi-capteurs par filtrage particulaire pour le recalage de navigation inertielle par corrélation de cartes", CIFRE, supervised by P. Del Moral, Dann Laneuville (NavalGroup) and Karim Dahia (ONERA)

11 Scientific production

11.1 Major publications

[1] M. Arnaudon and P. Del Moral. 'A second order analysis of McKean-Vlasov semigroups'. In: *Annals of Applied Probability* (2020). DOI: 10.1214/20-AAP1568. URL: https://hal.archives-ouvertes.fr/hal-02151808.

[2] B. Berthelot, E. Grivel, P. Legrand, J.-M. André and P. Mazoyer. 'Alternative Ways to Compare the Detendred Fluctuation Analysis and its Variants. Application to Visual Tunneling Detection'. In: *Digital Signal Processing* (2020). DOI: 10.1016/j.dsp.2020.102865. URL: https://hal.archives-ouvertes.fr/hal-02940122.

[3] I. Charlier, D. Paindaveine and J. Saracco. 'Multiple-output quantile regression through optimal quantization'. In: *Scandinavian Journal of Statistics* 47.1 (Feb. 2020), pp. 250–278. DOI: 10.1111/sjos.12426. URL: https://hal.archives-ouvertes.fr/hal-02429263.

[4] F. Dufour and A. Genadot. 'A Convex Programming Approach for Discrete-Time Markov Decision Processes under the Expected Total Reward Criterion'. In: *SIAM Journal on Control and Optimization* 58.4 (Jan. 2020), pp. 2535–2566. DOI: 10.1137/19M1255811. URL: https://hal.inria.fr/hal-03033727.

[5] M.-P. Ellies-Oury, J.-F. Hocquette, S. S. Chriki, A. Conanec, L. Farmer, M. Chavent and J. Saracco. 'Various Statistical Approaches to Assess and Predict Carcass and Meat Quality Traits'. In: *Foods* 9.4 (2020), p. 525. DOI: 10.3390/foods9040525. URL: https://hal.inrae.fr/hal-02570376.

11.2 Publications of the year

International journals

[6] J. Anselmi and F. Dufour. 'Power-of-d-Choices with Memory: Fluid Limit and Optimality'. In: *Mathematics of Operations Research* 45.3 (2020), pp. 862–888. URL: https://hal.archives-ouvertes.fr/hal-02394147.

[7] M. Arnaudon and P. Del Moral. 'A duality formula and a particle Gibbs sampler for continuous time Feynman-Kac measures on path spaces'. In: *Electronic Journal of Probability* (2020). DOI: 10.1214/20-EJP546. URL: https://hal.archives-ouvertes.fr/hal-01787257.

[8] M. Arnaudon and P. Del Moral. 'A second order analysis of McKean-Vlasov semigroups'. In: *Annals of Applied Probability* (2020). DOI: 10.1214/20-AAP1568. URL: https://hal.archives-ouvertes.fr/hal-02151808.

[9] B. Berthelot, E. Grivel, P. Legrand, J.-M. André and P. Mazoyer. 'Alternative Ways to Compare the Detendred Fluctuation Analysis and its Variants. Application to Visual Tunneling Detection'. In: *Digital Signal Processing* (2020). DOI: 10.1016/j.dsp.2020.102865. URL: https://hal.archives-ouvertes.fr/hal-02940122.

[10] A. N. Bishop, P. Del Moral and A. Niclas. 'A perturbation analysis of stochastic matrix Riccati diffusions'. In: *Annales de l'Institut Henri Poincaré. Section B. Calculs des Probabilités et Statistiques* (2020). DOI: 10.1214/19-AIHP987. URL: https://hal.inria.fr/hal-01593830.

[11] I. Charlier, D. Paindaveine and J. Saracco. 'Multiple-output quantile regression through optimal quantization'. In: *Scandinavian Journal of Statistics* 47.1 (18th Feb. 2020), pp. 250–278. DOI: 10.1111/sjos.12426. URL: https://hal.archives-ouvertes.fr/hal-02429263.

[12] M. Chavent, R. Genuer and J. Saracco. 'Combining clustering of variables and feature selection using random forests'. In: *Communications in Statistics - Simulation and Computation* 50.2 (11th Jan. 2021), pp. 426–445. DOI: 10.1080/03610918.2018.1563145. URL: https://hal.archives-ouvertes.fr/hal-02013631.

[13] P. Del Moral and S. Singh. 'A backward Itô–Ventzell formula with an application to stochastic interpolation'. In: *Comptes Rendus Mathématique* 358.7 (2020), pp. 881–886. DOI: 10.5802/crmath.110. URL: https://hal.inria.fr/hal-03122845.

[14] F. Dufour and A. Genadot. 'A Convex Programming Approach for Discrete-Time Markov Decision Processes under the Expected Total Reward Criterion'. In: *SIAM Journal on Control and Optimization* 58.4 (Jan. 2020), pp. 2535–2566. DOI: 10.1137/19M1255811. URL: https://hal.inria.fr/hal-03033727.

[15] F. Dufour and A. Genadot. 'On the Expected Total Reward with Unbounded Returns for Markov Decision Processes'. In: *Applied Mathematics and Optimization* 82.2 (2020), pp. 433–450. DOI: 10.1007/s00245-018-9533-6. URL: https://hal.inria.fr/hal-01953985.

[16] M.-P. Ellies-Oury, J.-F. Hocquette, S. S. Chriki, A. Conanec, L. Farmer, M. Chavent and J. Saracco. 'Various Statistical Approaches to Assess and Predict Carcass and Meat Quality Traits'. In: *Foods* 9.4 (2020), p. 525. DOI: 10.3390/foods9040525. URL: https://hal.inrae.fr/hal-02570376.

[17] S. Michel, D. Swingedouw, M. Chavent, P. Ortega, J. Mignot and M. Khodri. 'Reconstructing climatic modes of variability from proxy records using ClimIndRec version 1.0'. In: *Geoscientific Model Development* 13.2 (2020), pp. 841–858. DOI: 10.5194/gmd-13-841-2020. URL: https://hal.sorbonne-universite.fr/hal-02556996.

[18] N. Randriamihamison, N. Vialaneix and P. Neuvial. 'Applicability and Interpretability of Ward's Hierarchical Agglomerative Clustering With or Without Contiguity Constraints'. In: *Journal of Classification* (2020). DOI: 10.1007/s00357-020-09377-yâĂŃ. URL: https://hal.archives-ouvertes.fr/hal-02294847.

[19] M. Saleh, S.-M. Omar, E. Grivel and P. Legrand. 'A Variable Chirp Rate Stepped Frequency Linear Frequency Modulation Waveform Designed to Approximate Wideband Non-Linear Radar Waveforms'. In: *Digital Signal Processing* (2020). DOI: 10.1016/j.dsp.2020.102884. URL: https://hal.archives-ouvertes.fr/hal-02963775.

[20] L. Trujillo, U. Lopez and P. Legrand. 'SOAP: Semantic Outliers Automatic Preprocessing'. In: *Information Sciences* 526.81-101 (2020), p. 20. DOI: 10.1016/j.ins.2020.03.071. URL: https://hal.inria.fr/hal-02551161.

[21] E. Z-Flores, L. Trujillo, P. Legrand and F. Faïta-Aïnseba. 'EEG Feature Extraction Using Genetic Programming for the Classification of Mental States'. In: *Algorithms* 13.9 (Sept. 2020), p. 221. DOI: 10.3390/a13090221. URL: https://hal.inria.fr/hal-02943474.

International peer-reviewed conferences

[22] B. Berthelot, E. Grivel and P. Legrand. 'New variants of DFA based on loess and lowess methods: generalization of the detrending moving average'. In: ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing. Totonto, Canada, 2021. URL: https://hal.archives-ouvertes.fr/hal-03125981.

[23] B. Berthelot, E. Grivel, P. Legrand, J.-M. André and P. Mazoyer. 'Regularized Dfa To Study The Gaze Position Of An Airline Pilot'. In: EUSIPCO 2020 - 28th European Signal Processing Conference. Amsterdam / Virtual, Netherlands, 2020. URL: https://hal.archives-ouvertes.fr/hal-02716136.

[24] J. Bondu, E. Grivel, A. Giremus, P. Legrand, V. Corretja and M. Pommier. 'Sea Target Classification Based On An A Priori Motion Model'. In: EUSIPCO 2020 - 28th European Signal Processing Conference. Amsterdam / Virtual, Netherlands, 2020. DOI: 10.23919/Eusipco47968.2020.9287480. URL: https://hal.archives-ouvertes.fr/hal-02716100.

[25] U. Lopez, L. Trujillo, S. Silva, L. Vanneschi and P. Legrand. 'Unlabeled Multi-Target Regression with Genetic Programming'. In: GECCO 2020 - The Genetic and Evolutionary Computation Conference. Cancun, Mexico, 8th July 2020. URL: https://hal.inria.fr/hal-02551154.

Conferences without proceedings

[26] M. Chavent, J. Lacaille, A. Mourer and M. Olteanu. 'Sparse k-means for mixed data via group-sparse clustering'. In: ESANN 2020 - 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Vol. 978-2-87587-074-2. Bruges / Virtual, Belgium: https://www.esann.org/, 2nd Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-03130672.

[27] A. Conanec, M. Chavent, M. P. Ellies-Oury and J. Saracco. 'Une méthodologie computationnelle pour faire de l'optimisation multi-objectifs en élevage de précision'. In: JDS 2020 - 52èmes Journées de Statistique de la Société Française de Statistique. Nice, France, 25th May 2020. URL: https://hal.archives-ouvertes.fr/hal-03137866.

[28] B. Liquet and J. Saracco. 'BIG-SIR a sliced Inverse Regression approach for massive data'. In: CM-Statistics 2020 - 13th International Conference of the ERCIM WG on Computational and Methodological Statistic. Londres / Virtual, United Kingdom, 19th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03137853.

[29] A. Mourer, J. Lacaille, M. Olteanu and M. Chavent. 'Automatic Detection of Rare Observations During Production Tests Using Statistical Models'. In: PHM 2020 - Annual Conference of the PHM Society. Nashville, United States, 9th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03130682.

Scientific book chapters

[30] A. Gayte-Papon de Lameigné, P. Legrand and J. Lévy-Vehel. 'La modélisation de l'indemnisation du préjudice corporel'. In: *Le Big Data et le droit*. Feb. 2020, pp. 45–60. URL: https://hal.uca.fr/hal-02559847.

Doctoral dissertations and habilitation theses

[31] L. Labache. 'Elaboration Of Brain Network Atlases Underpinning Lateralized Cognitive Functions : Application To The Study Of Inter-individual Variability Of Language'. Université de Bordeaux, 23rd Oct. 2020. URL: https://tel.archives-ouvertes.fr/tel-03044068.

Reports & preprints

[32] A. N. Bishop and P. Del Moral. *On the Mathematical Theory of Ensemble (Linear-Gaussian) Kalman-Bucy Filtering*. 1st Dec. 2020. URL: https://hal.inria.fr/hal-03033604.

[33] M. Chavent and G. Chavent. *Optimal Projected Variance Group-Sparse Block PCA*. 29th Jan. 2021. URL: https://hal.inria.fr/hal-03125264.

[34] D. Crisan, P. Del Moral, A. Jasra and H. Ruzayqat. *Log-Normalization Constant Estimation using the Ensemble Kalman-Bucy Filter with Application to High-Dimensional Models*. 4th Feb. 2021. URL: https://hal.inria.fr/hal-03131613.

[35] B. Nguyen-Van-Yen, P. Del Moral and B. Cazelles. *Stochastic Epidemic Models inference and diagnosis with Poisson Random Measure Data Augmentation*. 1st Dec. 2020. URL: https://hal.inria.fr/hal-03033612.

[36] N. Randriamihamison, M. Chavent, S. Foissac, N. Vialaneix and P. Neuvial. *Analyse différentielle de données Hi-C via la classification ascendante hiérarchique sous contrainte de contiguïté*. Nice, France, 2020. URL: https://hal.archives-ouvertes.fr/hal-02892664.

11.3 Cited publications

[37] N. Duan and K.-C. Li. 'Slicing regression: a link-free regression method'. In: *Ann. Statist.* 19.2 (1991), pp. 505–530. DOI: 10.1214/aos/1176348109. URL: http://dx.doi.org/10.1214/aos/1176348109.

[38] R. Duda, P. Hart and D. Stork. *Pattern Classification*. John Wiley, 2001.

[39] K.-C. Li. 'Sliced inverse regression for dimension reduction'. In: *J. Amer. Statist. Assoc.* 86.414 (1991). With discussion and a rejoinder by the author, pp. 316–342.