RESEARCH CENTRE

Grenoble - Rhône-Alpes

IN PARTNERSHIP WITH:

CNRS, Institut polytechnique de Grenoble

2020
ACTIVITY REPORT

Project-Team
STATIFY

Bayesian and extreme value statistical models for structured and high dimensional data

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Contents

Project-Team STATIFY

Creation of the Project-Team: 2020 April 01

Keywords

Computer sciences and digital sciences

- A3.1.1. – Modeling, representation
- A3.1.4. – Uncertain data
- A3.3.2. – Data mining
- A3.3.3. – Big data analysis
- A3.4.1. – Supervised learning
- A3.4.2. – Unsupervised learning
- A3.4.4. – Optimization and learning
- A3.4.5. – Bayesian methods
- A3.4.7. – Kernel methods
- A5.3.3. – Pattern recognition
- A5.9.2. – Estimation, modeling
- A6.2. – Scientific computing, Numerical Analysis & Optimization
- A6.2.3. – Probabilistic methods
- A6.2.4. – Statistical methods
- A6.3. – Computation-data interaction
- A6.3.1. – Inverse problems
- A6.3.3. – Data processing
- A6.3.5. – Uncertainty Quantification
- A9.2. – Machine learning
- A9.3. – Signal analysis

Other research topics and application domains

- B1.2.1. – Understanding and simulation of the brain and the nervous system
- B2.6.1. – Brain imaging
- B3.3. – Geosciences
- B3.4.1. – Natural risks
- B3.4.2. – Industrial risks and waste
- B3.5. – Agronomy
- B5.1. – Factory of the future
- B9.5.6. – Data science
- B9.11.1. – Environmental risks

1 Team members, visitors, external collaborators

Research Scientists

- Florence Forbes [Team leader, Inria, Senior Researcher, from Apr 2020, HDR]
- Sophie Achard [CNRS, Senior Researcher, from Apr 2020, HDR]
- Julyan Arbel [Inria, Researcher, from Apr 2020, HDR]
- Stephane Girard [Inria, Senior Researcher, from Apr 2020, HDR]

Faculty Member

- Jean-Baptiste Durand [Institut polytechnique de Grenoble, Associate Professor, from Apr 2020]

Post-Doctoral Fellows

- Pascal Dkengne Sielenou [Inria, from Apr 2020]
- Antoine Usseglio-Carleve [Inria, from Apr 2020 until Sep 2020]
- Pierre Wolinski [Oxford, from Oct 2020]

PhD Students

- Karina Ashurbekova [Inria, from Apr 2020 until Jun 2020]
- Meryem Bousebata [Univ Grenoble Alpes, from Apr 2020]
- Fabien Boux [Univ Grenoble Alpes, from Apr 2020 until Aug 2020]
- Daria Bystrova [Univ Grenoble Alpes, from Apr 2020]
- Lucrezia Carboni [Univ Grenoble Alpes, from Oct 2020]
- Alexandre Constantin [Univ Grenoble Alpes, from Apr 2020]
- Benoit Kugler [Univ Grenoble Alpes, from Apr 2020]
- Hana Lbath [Univ Grenoble Alpes, from Oct 2020]
- Minh Tri Lê [Invensense, CIFRE, from Jun 2020]
- Theo Moins [Inria, from Oct 2020]
- Veronica Munoz Ramirez [Univ Grenoble Alpes, from Apr 2020 until Sep 2020]
- Giovanni Poggiato [Univ Grenoble Alpes, from Apr 2020]
- Mariia Vladimirova [Inria, from Apr 2020]

Technical Staff

- Fei Zheng [Inria, Engineer, from Apr 2020 until Sep 2020]

Interns and Apprentices

- Lucrezia Carboni [Univ Grenoble Alpes, from Apr 2020 until Jul 2020]
- Sami Djouadi [Univ Grenoble Alpes, from Apr 2020 until May 2020]
- Jiajie Li [Univ Grenoble Alpes, from Apr 2020 until Aug 2020]
- Tony Zheng [Inria, from Apr 2020 until Jul 2020]

Administrative Assistants

- Geraldine Christin [Inria, from Dec 2020]
- Marion Ponsot [Inria, from Apr 2020 until Oct 2020]

Visiting Scientists

- Mario Beraha [Politecnico di Milano, from Oct 2020]
- Jonathan El-Methni [Université de Paris, from Oct 2020]
- Trung Tin Nguyen [Univ de Caen Basse-Normandie, from Sep 2020]

External Collaborator

- Sami Djouadi [Univ Grenoble Alpes, from Jun 2020 until Jul 2020]

2 Overall objectives

The STATIFY team focuses on statistics. Statistics can be defined as a science of variation where the main question is how to acquire knowledge in the face of variation. In the past, statistics were seen as an opportunity to play in various backyards. Today, the statistician sees his own backyard invaded by data scientists, machine learners and other computer scientists of all kinds. Everyone wants to do data analysis and some (but not all) do it very well. Generally, data analysis algorithms and associated network architectures are empirically validated using domain-specific datasets and data challenges. While winning such challenges is certainly rewarding, statistical validation lies on more fundamentally grounded bases and raises interesting theoretical, algorithmic and practical insights. Statistical questions can be converted to probability questions by the use of probability models. Once certain assumptions about the mechanisms generating the data are made, statistical questions can be answered using probability theory. However, the proper formulation and checking of these probability models is just as important, or even more important, than the subsequent analysis of the problem using these models. The first question is then how to formulate and evaluate probabilistic models for the problem at hand. The second question is how to obtain answers after a certain model has been assumed. This latter task can be more a matter of applied probability theory, and in practice, contains optimization and numerical analysis.

The STATIFY team aims at bringing strengths, at a time when the number of solicitations received by statisticians increases considerably because of the successive waves of *big data, data science* and *deep learning*. The difficulty is to back up our approaches with reliable mathematics while what we have is often only empirical observations that we are not able to explain. Guiding data analysis with statistical justification is a challenge in itself. STATIFY has the ambition to play a role in this task and to provide answers to questions about the appropriate usage of statistics.

Often statistical assumptions do not hold. Under what conditions then can we use statistical methods to obtain reliable knowledge? These conditions are rarely the natural state of complex systems. The central motivation of STATIFY is to establish the conditions under which statistical assumptions and associated inference procedures approximately hold and become reliable.

However, as George Box said "Statisticians and artists both suffer from being too easily in love with their models". To moderate this risk, we choose to develop, in the team, expertise from different statistical domains to offer different solutions to attack a variety of problems. This is possible because these domains share the same mathematical food chain, from probability and measure theory to statistical modeling, inference and data analysis.

Our goal is to exploit methodological resources from statistics and machine learning to develop models that handle variability and that scale to high dimensional data while maintaining our ability to assess their correctness, typically the uncertainty associated with the provided solutions. To reach this goal, the team offers a unique range of expertise in statistics, combining probabilistic graphical models and mixture models to analyze structured data, Bayesian analysis to model knowledge and regularize ill-posed problems, non-parametric statistics, risk modeling and extreme value theory to face the lack, or impossibility, of precise modeling information and data. In the team, this expertise is organized to target five key challenges:

1. *Models for high dimensional, multimodal, heterogeneous data;*
2. *Spatial (structured) data science;*
3. *Scalable Bayesian models and procedures;*
4. *Understanding mathematical properties of statistical and machine learning methods;*
5. *The big problem of small data.*

The first two challenges address sources of complexity coming from data, namely, the fact that observations can be: 1) high dimensional, collected from multiple sensors in varying conditions *i.e.* multimodal and heterogeneous and 2) inter-dependent with a known structure between variables or with unknown interactions to be discovered. The other three challenges focus on providing reliable and interpretable models: 3) making the Bayesian approach scalable to handle large and complex data; 4) quantifying the information processing properties of machine learning methods and 5) allowing to draw reliable conclusions from datasets that are too small or not large enough to be used for training machine/deep learning methods.

These challenges rely on our four research axes:

1. *Models for graphs and networks;*
2. *Dimension reduction and latent variable modeling;*
3. *Bayesian modeling;*
4. *Modeling and quantifying extreme risk.*

In terms of applied work, we will target high-impact applications in neuroimaging, environmental and earth sciences.

3 Research program

3.1 Mixture models

Participants Jean-Baptiste Durand, Florence Forbes, Stephane Girard, Julyan Arbel, Daria Bystrova, Giovanni Poggiato, Fabien Boux, Veronica Munoz Ramirez, Benoit Kugler, Alexandre Constantin, Fei Zheng.

Keywords: Key-words: mixture of distributions, EM algorithm, missing data, conditional independence, statistical pattern recognition, clustering, unsupervised and partially supervised learning..

In a first approach, we consider statistical parametric models, θ being the parameter, possibly multi-dimensional, usually unknown and to be estimated. We consider cases where the data naturally divides into observed data $y = \{y_1, \ldots, y_n\}$ and unobserved or missing data $z = \{z_1, \ldots, z_n\}$. The missing data z_i represents for instance the memberships of one of a set of K alternative categories. The distribution of an observed y_i can be written as a finite mixture of distributions,

$$f(y_i;\theta) = \sum_{k=1}^{K} P(z_i = k;\theta) f(y_i \mid z_i;\theta) \,. \tag{1}$$

These models are interesting in that they may point out hidden variables responsible for most of the observed variability and so that the observed variables are *conditionally* independent. Their estimation is often difficult due to the missing data. The Expectation-Maximization (EM) algorithm is a general and now standard approach to maximization of the likelihood in missing data problems. It provides parameter estimation but also values for missing data.

Mixture models correspond to independent z_i's. They have been increasingly used in statistical pattern recognition. They enable a formal (model-based) approach to (unsupervised) clustering.

3.2 Markov models

Participants Jean-Baptiste Durand, Florence Forbes, Karina Ashurbekova, Julyan Arbel, Mariia Vladimirova.

Keywords: Key-words: graphical models, Markov properties, hidden Markov models, clustering, missing data, mixture of distributions, EM algorithm, image analysis, Bayesian inference..

Graphical modelling provides a diagrammatic representation of the dependency structure of a joint probability distribution, in the form of a network or graph depicting the local relations among variables. The graph can have directed or undirected links or edges between the nodes, which represent the individual variables. Associated with the graph are various Markov properties that specify how the graph encodes conditional independence assumptions.

It is the conditional independence assumptions that give graphical models their fundamental modular structure, enabling computation of globally interesting quantities from local specifications. In this way graphical models form an essential basis for our methodologies based on structures.

The graphs can be either directed, e.g. Bayesian Networks, or undirected, e.g. Markov Random Fields. The specificity of Markovian models is that the dependencies between the nodes are limited to the nearest neighbor nodes. The neighborhood definition can vary and be adapted to the problem of interest. When parts of the variables (nodes) are not observed or missing, we refer to these models as Hidden Markov Models (HMM). Hidden Markov chains or hidden Markov fields correspond to cases where the z_i's in (1) are distributed according to a Markov chain or a Markov field. They are a natural extension of mixture models. They are widely used in signal processing (speech recognition, genome sequence analysis) and in image processing (remote sensing, MRI, etc.). Such models are very flexible in practice and can naturally account for the phenomena to be studied.

Hidden Markov models are very useful in modelling spatial dependencies but these dependencies and the possible existence of hidden variables are also responsible for a typically large amount of computation. It follows that the statistical analysis may not be straightforward. Typical issues are related to the neighborhood structure to be chosen when not dictated by the context and the possible high dimensionality of the observations. This also requires a good understanding of the role of each parameter and methods to tune them depending on the goal in mind. Regarding estimation algorithms, they correspond to an energy minimization problem which is NP-hard and usually performed through approximation. We focus on a certain type of methods based on variational approximations and propose effective algorithms which show good performance in practice and for which we also study theoretical properties. We also propose some tools for model selection. Eventually we investigate ways to extend the standard Hidden Markov Field model to increase its modelling power.

3.3 Functional Inference, semi- and non-parametric methods

Participants Julyan Arbel, Daria Bystrova, Giovanni Poggiato, Stephane Girard, Florence Forbes, Antoine Usseglio Carleve, Pascal Dkengne Sielenou, Meryem Bousebata, Sophie Achard
.

Keywords: Key-words: dimension reduction, extreme value analysis, functional estimation..

We also consider methods which do not assume a parametric model. The approaches are non-parametric in the sense that they do not require the assumption of a prior model on the unknown quantities. This property is important since, for image applications for instance, it is very difficult to introduce sufficiently general parametric models because of the wide variety of image contents. Projection methods are then a way to decompose the unknown quantity on a set of functions (*e.g.* wavelets). Kernel methods which rely on smoothing the data using a set of kernels (usually probability distributions) are other examples. Relationships exist between these methods and learning techniques using Support

Vector Machine (SVM) as this appears in the context of *level-sets estimation* (see section 3.5). Such non-parametric methods have become the cornerstone when dealing with functional data [66]. This is the case, for instance, when observations are curves. They enable us to model the data without a discretization step. More generally, these techniques are of great use for *dimension reduction* purposes (section 3.6). They enable reduction of the dimension of the functional or multivariate data without assumptions on the observations distribution. Semi-parametric methods refer to methods that include both parametric and non-parametric aspects. Examples include the Sliced Inverse Regression (SIR) method [68] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [65], which is based on the modelling of distribution tails (see section 3.4). It differs from traditional statistics which focuses on the central part of distributions, *i.e.* on the most probable events. Extreme value theory shows that distribution tails can be modelled by both a functional part and a real parameter, the extreme value index.

3.4 Modelling extremal events

Extreme value theory is a branch of statistics dealing with the extreme deviations from the bulk of probability distributions. More specifically, it focuses on the limiting distributions for the minimum or the maximum of a large collection of random observations from the same arbitrary distribution. Let $X_{1,n} \leq \ldots \leq X_{n,n}$ denote n ordered observations from a random variable X representing some quantity of interest. A p_n-quantile of X is the value x_{p_n} such that the probability that X is greater than x_{p_n} is p_n, *i.e.* $P(X > x_{p_n}) = p_n$. When $p_n < 1/n$, such a quantile is said to be extreme since it is usually greater than the maximum observation $X_{n,n}$.

To estimate such quantiles therefore requires dedicated methods to extrapolate information beyond the observed values of X. Those methods are based on Extreme value theory. This kind of issue appeared in hydrology. One objective was to assess risk for highly unusual events, such as 100-year floods, starting from flows measured over 50 years. To this end, semi-parametric models of the tail are considered:

$$P(X > x) = x^{-1/\theta}\ell(x),\ x > x_0 > 0, \tag{2}$$

where both the extreme-value index $\theta > 0$ and the function $\ell(x)$ are unknown. The function ℓ is a slowly varying function *i.e.* such that

$$\frac{\ell(tx)}{\ell(x)} \to 1 \text{ as } x \to \infty \tag{3}$$

for all $t > 0$. The function $\ell(x)$ acts as a nuisance parameter which yields a bias in the classical extreme-value estimators developed so far. Such models are often referred to as heavy-tail models since the probability of extreme events decreases at a polynomial rate to zero. It may be necessary to refine the model (2,3) by specifying a precise rate of convergence in (3). To this end, a second order condition is introduced involving an additional parameter $\rho \leq 0$. The larger ρ is, the slower the convergence in (3) and the more difficult the estimation of extreme quantiles.

More generally, the problems that we address are part of the risk management theory. For instance, in reliability, the distributions of interest are included in a semi-parametric family whose tails are decreasing exponentially fast. These so-called Weibull-tail distributions [10] are defined by their survival distribution function:

$$P(X > x) = \exp\{-x^{\theta}\ell(x)\},\ x > x_0 > 0. \tag{4}$$

Gaussian, gamma, exponential and Weibull distributions, among others, are included in this family. An important part of our work consists in establishing links between models (2) and (4) in order to propose new estimation methods. We also consider the case where the observations were recorded with a covariate information. In this case, the extreme-value index and the p_n-quantile are functions of the covariate. We propose estimators of these functions by using moving window approaches, nearest neighbor methods, or kernel estimators.

3.5 Level sets estimation

Level sets estimation is a recurrent problem in statistics which is linked to outlier detection. In biology, one is interested in estimating reference curves, that is to say curves which bound 90% (for example) of the population. Points outside this bound are considered as outliers compared to the reference population. Level sets estimation can be looked at as a conditional quantile estimation problem which benefits from a non-parametric statistical framework. In particular, boundary estimation, arising in image segmentation as well as in supervised learning, is interpreted as an extreme level set estimation problem. Level sets estimation can also be formulated as a linear programming problem. In this context, estimates are sparse since they involve only a small fraction of the dataset, called the set of support vectors.

3.6 Dimension reduction

Our work on high dimensional data requires that we face the curse of dimensionality phenomenon. Indeed, the modelling of high dimensional data requires complex models and thus the estimation of high number of parameters compared to the sample size. In this framework, dimension reduction methods aim at replacing the original variables by a small number of linear combinations with as small as a possible loss of information. Principal Component Analysis (PCA) is the most widely used method to reduce dimension in data. However, standard linear PCA can be quite inefficient on image data where even simple image distortions can lead to highly non-linear data. Two directions are investigated. First, non-linear PCAs can be proposed, leading to semi-parametric dimension reduction methods [67]. Another field of investigation is to take into account the application goal in the dimension reduction step. One of our approaches is therefore to develop new Gaussian models of high dimensional data for parametric inference [64]. Such models can then be used in a Mixtures or Markov framework for classification purposes. Another approach consists in combining dimension reduction, regularization techniques, and regression techniques to improve the Sliced Inverse Regression method [68].

4 Application domains

4.1 Image Analysis

Participants Veronica Munoz Ramirez, Florence Forbes, Stephane Girard, Fabien Boux, Benoit Kugler, Alexandre Constantin.

As regards applications, several areas of image analysis can be covered using the tools developed in the team. More specifically, in collaboration with team PERCEPTION, we address various issues in computer vision involving Bayesian modelling and probabilistic clustering techniques. Other applications in medical imaging are natural. We work more specifically on MRI and functional MRI data, in collaboration with the Grenoble Institute of Neuroscience (GIN). We also consider other statistical 2D fields coming from other domains such as remote sensing, in collaboration with the Institut de Planétologie et d'Astrophysique de Grenoble (IPAG) and the Centre National d'Etudes Spatiales (CNES). In this context, we worked on hyperspectral and/or multitemporal images. In the context of the "pole de competivité" project I-VP, we worked of images of PC Boards.

4.2 Biology, Environment and Medicine

Participants Florence Forbes, Stephane Girard, Jean-Baptiste Durand, Julyan Arbel, Sophie Achard, Karina Ashurbekova, Fabien Boux, Veronica Munoz Ramirez, Fei Zheng.

A third domain of applications concerns biology and medicine. We considered the use of mixture models to identify biomakers. We also investigated statistical tools for the analysis of fluorescence signals in molecular biology. Applications in neurosciences are also considered. In the environmental domain, we considered the modelling of high-impact weather events and the use of hyperspectral data as a new tool for quantitative ecology.

-

5 Highlights of the year

The new Statify team has been officially created on April 1, 2020.

5.1 New projects

- A new ANR France/USA project entitled Q-FunC has been founded for 4 years 2020-2023. It is coordinated by Sophie Achard and involves Statify, Grenoble Institute of Neuroscience and University of California Santa-Barbara. The goal is to study spatiotemporal statistical models for quantification and estimation of functional connectivity. The PhD of Hana Lbath has started in this context.
- Sophie Achard is also the co-pi of a new MIAI chair for 2020-2023. The Statify part is to investigate graph neural networks for brain connectivity exploration. The PhD of Lucrezia Carboni with Michel Dojat from GIN has started in this context.
- A new 3-year project and contract coordinated by Stephane Girard has been signed with EDF. The goal is to study the possibility to incorporate Bayesian statistic techniques into extreme value theory. The PhD of Theo Moins has started in this context.
- Stephane Girard is also the PI of a new contract with Valeo on the use of extreme value theory in the context of autonomous cars.
- A 2-year IDEX project coordinated by Julyan Arbel has been founded in collaboration with Judith Rousseau at University of Oxford. The post-doc of Pierre Wolinski on Bayesian deep learning is in this context.
- A three-year CIFRE contract has been signed in 2020 with TDK InvenSense for Ph.D thesis of Minh Tri Lê, co-advised by Julyan Arbel with Etienne De Foras. The topic deals with constrained signal processing using deep neural networks.

5.2 New responsabilities

Sophie Achard has been elected in November 2020 as the new director of the MSTIC pole at UGA.

6 New results

6.1 Mixture models

6.1.1 Global implicit function theorems and the online EM algorithm

Participants Florence Forbes.

Joint work with: Hien Nguyen, La Trobe University Melbourne Australia.

The expectation–maximisation (EM) algorithm is an important tool for statistical computation. Due to the changing nature of data, online and mini-batch variants of EM and EM-like algorithms have become increasingly popular. The consistency of the estimator sequences that are produced by these EM variants often rely on an assumption regarding the continuous differentiability of a parameter update function. In many cases, the parameter update function is often not in closed form and may only be defined implicitly, which makes the verification of the continuous differentiability property difficult. We demonstrate how a global implicit function theorem can be used to verify such properties in the cases of finite mixtures of distributions in the exponential family and more generally when the component specific distribution admits a data augmentation scheme in the exponential family. We demonstrate the use of such a theorem in the case of mixtures of beta distributions, gamma distributions, fully-visible Boltzmann machines and

Student distributions. Via numerical simulations, we provide empirical evidence towards the consistency of the online EM algorithm parameter estimates in such cases. Details can be found in [62].

6.1.2 Fast Bayesian Inversion for high dimensional inverse problems

Participants Florence Forbes, Benoit Kugler.

Joint work with: Sylvain Douté from Institut de Planétologie et d'Astrophysique de Grenoble (IPAG).

We investigated the use of learning approaches to handle Bayesian inverse problems in a computationally efficient way when the signals to be inverted present a moderately high number of dimensions and are in large number. We proposed a tractable inverse regression approach which has the advantage to produce full probability distributions as approximations of the target posterior distributions. In addition to provide confidence indices on the predictions, these distributions allow a better exploration of inverse problems when multiple equivalent solutions exist. We then showed how these distributions could be used for further refined predictions using importance sampling, while also providing a way to carry out uncertainty level estimation if necessary. The relevance of the proposed approach was illustrated both on simulated and real data in the context of a physical model inversion in planetary remote sensing. The approach showed interesting capabilities both in terms of computational efficiency and multimodal inference. Details can be found in [60].

6.1.3 Bayesian inverse regression for vascular magnetic resonance fingerprinting

Participants Florence Forbes, Fabien Boux, Julyan Arbel.

Joint work with: Emmanuel Barbier from Grenoble Institute of Neuroscience.

Standard parameter estimation from vascular magnetic resonance fingerprinting (MRF) data is based on matching the MRF signals to their best counterparts in a grid of coupled simulated signals and parameters, referred to as a dictionary. To reach a good accuracy, the matching requires an informative dictionary whose cost, in terms of design, storage and exploration, is rapidly prohibitive for even moderate numbers of parameters. In this work, we propose an alternative dictionary-based statistical learning (DB-SL) approach made of three steps: 1) a quasi-random sampling strategy to produce efficiently an informative dictionary, 2) an inverse statistical regression model to learn from the dictionary a correspondence between fingerprints and parameters, and 3) the use of this mapping to provide both parameter estimates and their confidence indices. The proposed DB-SL approach is compared to both the standard dictionary-based matching (DBM) method and to a dictionary-based deep learning (DB-DL) method. Performance is illustrated first on synthetic signals including scalable and standard MRF signals with spatial undersampling noise. Then, vascular MRF signals are considered both through simulations and real data acquired in tumor bearing rats. Overall, the two learning methods yield more accurate parameter estimates than matching and to a range not limited to the dictionary boundaries. DB-SL in particular resists to higher noise levels and provides in addition confidence indices on the estimates at no additional cost. DB-SL appears as a promising method to reduce simulation needs and computational requirements, while modeling sources of uncertainty and providing both accurate and interpretable results. More details can be found in [18].

6.1.4 Unannounced Meal Detection for Artificial Pancreas Systems Using Extended Isolation Forest.

Participants Florence Forbes, Fei Zheng.

Joint work with: Stéphane Bonnet from CEA Leti.

This study aims at developing an unannounced meal detection method for artificial pancreas, based on a recent extension of Isolation Forest. The proposed method makes use of features accounting for individual Continuous Glucose Monitoring (CGM) profiles and benefits from a two-threshold decision rule detection. The advantage of using Extended Isolation Forest (EIF) instead of the standard one is supported by experiments on data from virtual diabetic patients, showing good detection accuracy with acceptable detection delays.

6.1.5 Dirichlet process mixtures under affine transformations of the data

Participants Julyan Arbel.

Joint work with: Riccardo Corradin and Bernardo Nipoti from Milano Bicocca, Italy.

Location-scale Dirichlet process mixtures of Gaussians (DPM-G) have proved extremely useful in dealing with density estimation and clustering problems in a wide range of domains. Motivated by an astronomical application, in this work we address the robustness of DPM-G models to affine transformations of the data, a natural requirement for any sensible statistical method for density estimation. In [14], we first devise a coherent prior specification of the model which makes posterior inference invariant with respect to affine transformation of the data. Second, we formalize the notion of asymptotic robustness under data transformation and show that mild assumptions on the true data generating process are sufficient to ensure that DPM-G models feature such a property. As a by-product, we derive weaker assumptions than those provided in the literature for ensuring posterior consistency of Dirichlet process mixtures, which could reveal of independent interest. Our investigation is supported by an extensive simulation study and illustrated by the analysis of an astronomical dataset consisting of physical measurements of stars in the field of the globular cluster NGC 2419.

6.1.6 Approximate Bayesian computation with surrogate posteriors

Participants Julyan Arbel, Florence Forbes.

Joint work with: Hien Nguyen, La Trobe University Melbourne Australia and Trung Tin Nguyen from University Caen Normandy.

A key ingredient in approximate Bayesian computation (ABC) procedures is the choice of a discrepancy that describes how different the simulated and observed data are, often based on a set of summary statistics when the data cannot be compared directly. Unless discrepancies and summaries are available from experts or prior knowledge, which seldom occurs, they have to be chosen and this can affect the approximations. Their choice is an active research topic, which has mainly considered data discrepancies requiring samples of observations or distances between summary statistics, to date. In this work, we introduce a preliminary learning step in which surrogate posteriors are built from finite Gaussian mixtures using an inverse regression approach. These surrogate posteriors are then used in place of summary statistics and compared using metrics between distributions in place of data discrepancies. Two such metrics are investigated, a standard L_2 distance and an optimal transport-based distance. The whole procedure can be seen as an extension of the semi-automatic ABC framework to functional summary statistics. The resulting ABC quasi-posterior distribution is shown to converge to the true one, under standard conditions. Performance is illustrated on both synthetic and real data sets, where it is shown that our approach is particularly useful when the posterior is multimodal. Details can be found in [55].

6.1.7 Joint supervised classification and reconstruction of irregularly sampled satellite image times series

Participants Alexandre Constantin, Stephane Girard.

Joint work with: Mathieu Fauvel, INRAE

Recent satellite missions have led to a huge amount of earth observation data, most of them being freely available. In such a context, satellite image time series have been used to study land use and land cover information. However, optical time series, like Sentinel-2 or Landsat ones, are provided with an irregular time sampling for different spatial locations, and images may contain clouds and shadows. Thus, pre-processing techniques are usually required to properly classify such data. The proposed approach is able to deal with irregular temporal sampling and missing data directly in the classification process. It is based on Gaussian processes and allows to perform jointly the classification of the pixel labels as well as the reconstruction of the pixel time series. The method complexity scales linearly with the number of pixels, making it amenable in large scale scenarios. Experimental classification and reconstruction results show that the method does not compete yet with state of the art classifiers but yields reconstructions that are robust with respect to the presence of undetected clouds or shadows and does not require any temporal preprocessing [52].

6.2 Semi and non-parametric methods

6.2.1 Subtle anomaly detection in MRI brain scans: Application to biomarkers extraction in patients with *de novo* Parkinson's disease

Participants Florence Forbes, Veronica Munoz Ramirez, Virgilio Kmetzsch Rosa E Silva.

Joint work with: Michel Dojat from Grenoble Institute of Neuroscience and Elena Mora from CHUGA.

With the advent of recent deep learning techniques, computerized methods for automatic lesion segmentation have reached performances comparable to those of medical practitioners. However, little attention has been paid to the detection of subtle physiological changes caused by evolutive pathologies such as neurodegenerative diseases. In this work, we investigated the ability of deep learning models to detect anomalies in magnetic resonance imaging (MRI) brain scans of recently diagnosed and untreated (*de novo*) patients with Parkinson's disease (PD). We evaluated two families of auto-encoders, fully convolutional and variational auto-encoders. The models were trained with diffusion tensor imaging (DTI) parameter maps of healthy controls. Then, reconstruction errors computed by the models in different brain regions allowed to classify controls and patients with ROC AUC up to 0.81. Moreover, the white matter and the subcortical structures, particularly the substantia nigra, were identified as the regions the most impacted by the disease, in accordance with the physio-pathology of PD. Our results suggest that deep learning-based anomaly detection models, even trained on a moderate number of images, are promising tools for extracting robust neuroimaging biomarkers of PD. Interestingly, such models can be seamlessly extended with additional quantitative MRI parameters and could provide new knowledge about the physio-pathology of neuro-degenerative diseases.

6.2.2 Estimation of extreme risk measures

Participants Stephane Girard, Antoine Usseglio Carleve.

Joint work with: A. Daouia (Univ. Toulouse), L. Gardes (Univ. Strasbourg) and G. Stupfler (Ensai).

One of the most popular risk measures is the Value-at-Risk (VaR) introduced in the 1990's. In statistical terms, the VaR at level $\alpha \in (0,1)$ corresponds to the upper α-quantile of the loss distribution. The Value-at-Risk however suffers from several weaknesses. First, it provides us only with a pointwise information: VaR(α) does not take into consideration what the loss will be beyond this quantile. Second, random loss variables with light-tailed distributions or heavy-tailed distributions may have the same Value-at-Risk. Finally, Value-at-Risk is not a coherent risk measure since it is not subadditive in general. A first coherent alternative risk measure is the Conditional Tail Expectation (CTE), also known as Tail-Value-at-Risk, Tail Conditional Expectation or Expected Shortfall in case of a continuous loss distribution. The CTE is defined as the expected loss given that the loss lies above the upper α-quantile of the loss distribution. This risk measure thus takes into account the whole information contained in the upper tail of the distribution.

However, the asymptotic normality of the empirical CTE estimator requires that the underlying distribution possess a finite variance; this can be a strong restriction in heavy-tailed models which constitute the favoured class of models in actuarial and financial applications. One possible solution in very heavy-tailed models where this assumption fails could be to use the more robust Median Shortfall, but this quantity is actually just a quantile, which therefore only gives information about the frequency of a tail event and not about its typical magnitude. In [25], we construct a synthetic class of tail L_p-medians, which encompasses the Median Shortfall (for $p=1$) and Conditional Tail Expectation (for $p=2$). We show that, for $1<p<2$, a tail L_p-median always takes into account both the frequency and magnitude of tail events, and its empirical estimator is, within the range of the data, asymptotically normal under a condition weaker than a finite variance. We extrapolate this estimator, along with another technique, to proper extreme levels using the heavy-tailed framework. The estimators are showcased on a simulation study and on a set of real fire insurance data showing evidence of a very heavy right tail.

Risk measures of a financial position are, from an empirical point of view, mainly based on quantiles. Replacing quantiles with their least squares analogues, called expectiles, has recently received increasing attention. The novel expectile-based risk measures satisfy all coherence requirements. We revisit their extreme value estimation for heavy-tailed distributions. First, we estimate the underlying tail index via weighted combinations of top order statistics and asymmetric least squares estimates. The resulting expectHill estimators are then used as the basis for estimating tail expectiles and Expected Shortfall. The asymptotic theory of the proposed estimators is provided, along with numerical simulations and applications to actuarial and financial data [22].

The estimation of expectiles typically requires to consider non-explicit asymmetric least squares estimates rather than the traditional order statistics used for quantile estimation. This makes the study of the tail expectile process a lot harder than that of the standard tail quantile process. Under the challenging model of heavy-tailed distributions, we derive joint weighted Gaussian approximations of the tail empirical expectile and quantile processes. We then use this powerful result to introduce and study new estimators of extreme expectiles and the standard quantile-based expected shortfall, as well as a novel expectile-based form of expected shortfall. Our estimators are built on general weighted combinations of both top order statistics and asymmetric least squares estimates. Some numerical simulations and applications to actuarial and financial data are provided [23].

Currently available estimators of extreme expectiles are typically biased and hence may show poor finite-sample performance even in fairly large samples. In [59], we focus on the construction of bias-reduced extreme expectile estimators for heavy-tailed distributions. The rationale for our construction hinges on a careful investigation of the asymptotic proportionality relationship between extreme expectiles and their quantile counterparts, as well as of the extrapolation formula motivated by the heavy-tailed context. We accurately quantify and estimate the bias incurred by the use of these relationships when constructing extreme expectile estimators. This motivates the introduction of a class of bias-reduced estimators whose asymptotic properties are rigorously shown, and whose finite-sample properties are assessed on a simulation study and three samples of real data from economics, insurance and finance. The results are submitted for publication.

6.2.3 Conditional extremal events

Participants Stephane Girard, Antoine Usseglio Carleve.

Joint work with: G. Stupfler (Ensai), A. Ahmad, E. Deme and A. Diop (Université Gaston Berger, Sénégal).

The goal of the PhD thesis of Aboubacrene Ag Ahmad is to contribute to the development of theoretical and algorithmic models to tackle conditional extreme value analysis, *ie* the situation where some covariate information X is recorded simultaneously with a quantity of interest Y. In such a case, extreme quantiles and expectiles are functions of the covariate. In [11], we consider a location-scale model for conditional heavy-tailed distributions when the covariate is deterministic. First, nonparametric estimators of the location and scale functions are introduced. Second, an estimator of the conditional extreme-value index is derived. The asymptotic properties of the estimators are established under mild assumptions and their finite sample properties are illustrated both on simulated and real data.

As explained in Paragraph 6.2.2, expectiles have recently started to be considered as serious candidates to become standard tools in actuarial and financial risk management. However, expectiles and their sample versions do not benefit from a simple explicit form, making their analysis significantly harder than that of quantiles and order statistics. This difficulty is compounded when one wishes to integrate auxiliary information about the phenomenon of interest through a finite-dimensional covariate, in which case the problem becomes the estimation of conditional expectiles. In [26], we exploit the fact that the expectiles of a distribution F are in fact the quantiles of another distribution E explicitly linked to F, in order to construct nonparametric kernel estimators of extreme conditional expectiles. We analyze the asymptotic properties of our estimators in the context of conditional heavy-tailed distributions. Applications to simulated data and real insurance data are provided. The extension to functional covariates is investigated in [58] and submitted for publication.

In [57], we build a general theory for the estimation of extreme conditional expectiles in heteroscedastic regression models with heavy-tailed noise. Our approach is supported by general results of independent interest on residual-based extreme value estimators in heavy-tailed regression models, and is intended to cope with covariates having a large but fixed dimension. We demonstrate how our results can be applied to a wide class of important examples, among which linear models, single-index models as well as ARMA and GARCH time series models. Our estimators are showcased on a numerical simulation study and on real sets of actuarial and financial data. The results are submitted for publication.

6.2.4 Dimension reduction for extremes

Participants Meryem Bousebata, Stephane Girard.

Joint work with: G. Enjolras (CERAG). In the context of the PhD thesis of Meryem Bousebata, we propose

a new approach, called Extreme-PLS, for dimension reduction in regression and adapted to distribution tails. The objective is to find linear combinations of predictors that best explain the extreme values of the response variable in a non-linear inverse regression model. The asymptotic normality of the Extreme-PLS estimator is established in the single-index framework and under mild assumptions. The performance of the method is assessed on simulated data. A statistical analysis of French farm income data, considering extreme cereal yields, is provided as an illustration. The results are submitted for publication [49].

6.2.5 Estimation of the variability in the distribution tail

Participants Stephane Girard.

Joint work with: L. Gardes (Univ. Strasbourg).

We propose a new measure of variability in the tail of a distribution by applying a Box-Cox transformation of parameter $p \geq 0$ to the tail-Gini functional. It is shown that the so-called Box-Cox Tail Gini Variability measure is a valid variability measure whose condition of existence may be as weak as necessary thanks to the tuning parameter p. The tail behaviour of the measure is investigated under a general extreme-value condition on the distribution tail. We then show how to estimate the Box-Cox Tail Gini Variability measure within the range of the data. These methods provide us with basic estimators that are then extrapolated using the extreme-value assumption to estimate the variability in the very far tails. The finite sample behavior of the estimators is illustrated both on simulated and real data. This work is published in [24].

6.2.6 Extrapolation limits associated with extreme-value methods

Participants Stephane Girard.

Joint work with: L. Gardes (Univ. Strasbourg) and A. Dutfoy (EDF R&D).

In [13], we investigate the asymptotic behavior of the (relative) extrapolation error associated with some estimators of extreme quantiles based on extreme-value theory. It is shown that the extrapolation error can be interpreted as the remainder of a first order Taylor expansion. Necessary and sufficient conditions are then provided such that this error tends to zero as the sample size increases. Interestingly, in case of the so-called Exponential Tail estimator, these conditions lead to a subdivision of Gumbel maximum domain of attraction into three subsets. In contrast, the extrapolation error associated with Weissman estimator has a common behavior over the whole Fréchet maximum domain of attraction. First order equivalents of the extrapolation error are then derived and their accuracy is illustrated numerically.

In [12], we propose a new estimator for extreme quantiles under the log-generalized Weibull-tail model, introduced by Cees de Valk. This model relies on a new regular variation condition which, in some situations, permits to extrapolate further into the tails than the classical assumption in extreme-value theory. The asymptotic normality of the estimator is established and its finite sample properties are illustrated both on simulated and real datasets.

6.2.7 Approximations of Bayesian nonparametric models

Participant Julyan Arbel, Daria Bystrova.

Joint work with: Stefano Favaro from Collegio Carlo Alberto, Turin, Italy, Guillaume Kon Kam King and François Deslandes from MaIAGE - Mathématiques et Informatique Appliquées du Génome à l'Environnement (INRAE Jouy-En-Josas)

In this work, we approximate predictive probabilities of Gibbs-type random probability measures, or Gibbs-type priors, which are arguably the most "natural" generalization of the celebrated Dirichlet prior. Among them the Pitman–Yor process certainly stands out for the mathematical tractability and interpretability of its predictive probabilities, which made it the natural candidate in several applications. Given a sample of size n, in this paper we show that the predictive probabilities of any Gibbs-type prior admit a large n approximation, with an error term vanishing as $o(1/n)$, which maintains the same desirable features as the predictive probabilities of the Pitman–Yor process.

In [37], we study the prior distribution induced on the number of clusters, which is key for prior specification and calibration. However, evaluating this prior is infamously difficult even for moderate sample size. We evaluate several statistical approximations to the prior distribution on the number of clusters for Gibbs-type processes, a class including the Pitman-Yor process and the normalized generalized gamma process. We introduce a new approximation based on the predictive distribution of Gibbs-type process, which compares favourably with the existing methods. We thoroughly discuss the limitations of these various approximations by comparing them against an exact implementation of the prior distribution of the number of clusters.

We prove a monotonicity property of the Hurwitz zeta function which, in turn, translates into a chain of inequalities for polygamma functions of different orders. We provide a probabilistic interpretation of our result by exploiting a connection between Hurwitz zeta function and the cumulants of the exponential-beta distribution.

6.2.8 Concentration inequalities

Participant Julyan Arbel, Stéphane Girard, Mariia Vladimirova.

Joint work with: Olivier Marchal from Université Jean Monnet and Hien Nguyen from La Trobe University Melbourne Australia.

In this work, we investigate the sub-Gaussian property for almost surely bounded random variables. If sub-Gaussianity per se is de facto ensured by the bounded support of said random variables, then exciting research avenues remain open. Among these questions is how to characterize the optimal sub-Gaussian proxy variance? Another question is how to characterize strict sub-Gaussianity, defined by a proxy variance equal to the (standard) variance? We address the questions in proposing conditions based on the study of functions variations. A particular focus is given to the relationship between strict sub-Gaussianity and symmetry of the distribution. In particular, we demonstrate that symmetry is neither sufficient nor necessary for strict sub-Gaussianity. In contrast, simple necessary conditions on the one hand, and simple sufficient conditions on the other hand, for strict sub-Gaussianity are provided. These results are illustrated via various applications to a number of bounded random variables, including Bernoulli, beta, binomial, uniform, Kumaraswamy, and triangular distributions.

In [34], we propose the notion of sub-Weibull distributions, which are characterised by tails lighter than (or equally light as) the right tail of a Weibull distribution. This novel class generalises the sub-Gaussian and sub-Exponential families to potentially heavier-tailed distributions. Sub-Weibull distributions are parameterized by a positive tail index θ and reduce to sub-Gaussian distributions for $\theta = 1/2$ and to sub-Exponential distributions for $\theta = 1$. A characterisation of the sub-Weibull property based on moments and on the moment generating function is provided and properties of the class are studied. An estimation procedure for the tail parameter is proposed and is applied to an example stemming from Bayesian deep learning.

6.2.9 Applications of semi and non-parametric methods in ecology and genomics

Participant Julyan Arbel, Daria Bystrova, Giovanni Poggiato.

Joint work with: Florian Privé and Bjarni Vilhjálmsson from National Center for Register-Based Research (Aarhus, Denmark), Billur Bektaş and Wilfried Thuiller from LECA - Laboratoire d'Ecologie Alpine, James S Clark from Nicholas School of the Environment, Duke University, USA, Alessandra Guglielmi from POLIMI - Dipartimento di Matematica - POLIMI, Politecnico di Milano.

In [21], we investigate modelling species distributions over space and time which is one of the major research topics in both ecology and conservation biology. Joint Species Distribution models (JSDMs) have recently been introduced as a tool to better model community data, by inferring a residual covariance matrix between species, after accounting for species' response to the environment. However, these models are computationally demanding, even when latent factors, a common tool for dimension reduction, are used. To address this issue, previous research proposed to use a Dirichlet process, a Bayesian nonparametric prior, to further reduce model dimension by clustering species in the residual covariance matrix. Here, we built on this approach to include a prior knowledge on the potential number of clusters, and instead used a Pitman-Yor process to address some critical limitations of the Dirichlet process. We therefore propose a framework that includes prior knowledge in the residual covariance matrix, providing a tool to analyze clusters of species that share the same residual associations with respect to other species. We applied our methodology to a case study of plant communities in a protected area of the French Alps (the Bauges Regional Park), and demonstrated that our extensions improve dimension reduction and reveal additional information from the residual covariance matrix, notably showing how the estimated clusters are compatible with plant traits, endorsing their importance in shaping communities.

In [31], we investigate modelling polygenic scores which have become a central tool in human genetics research. LDpred is a popular method for deriving polygenic scores based on summary statistics and a matrix of correlation between genetic variants. However, LDpred has limitations that may reduce its predictive performance. Here we present LDpred2, a new version of LDpred that addresses these issues. We also provide two new options in LDpred2: a "sparse" option that can learn effects that are exactly 0, and an "auto" option that directly learns the two LDpred parameters from data. We benchmark predictive performance of LDpred2 against the previous version on simulated and real data, demonstrating substantial improvements in robustness and predictive accuracy compared to LDpred1. We then show that LDpred2 also outperforms other polygenic score methods recently developed, with a mean AUC over the 8 real traits analyzed here of 65.1%, compared to 63.8% for lassosum, 62.9% for PRS-CS and 61.5% for SBayesR. Note that LDpred2 provides more accurate polygenic scores when run genome-wide, instead of per chromosome. LDpred2 is implemented in R package bigsnpr.

6.3 Graphical and Markov models

6.3.1 Optimal shrinkage for robust covariance matrix estimators in a small sample size setting

Participants Sophie Achard, Karina Ashurbekova, Florence Forbes, Antoine Usseglio Carleve.

When estimating covariance matrices, traditional sample covariance-based estimators are straightforward but suffer from two main issues: 1) a lack of robustness, which occurs as soon as the samples do not come from a Gaussian distribution or are contaminated with outliers and 2) a lack of data when the number of parameters to estimate is too large compared to the number of available observations, which occurs as soon as the covariance matrix dimension is greater than the sample size. The first issue can be handled by assuming samples are drawn from a heavy-tailed distribution, at the cost of more complex derivations, while the second issue can be addressed by shrinkage with the difficulty of choosing the appropriate level of regularization. In this work [48] we offer both a tractable and optimal framework based on shrinked likelihood-based M-estimators. First, a closed-form expression is provided for a regularized covariance matrix estimator with an optimal shrinkage coefficient for any sample distribution in the elliptical family. Then, a complete inference procedure is proposed which can also handle both unknown mean and tail parameter, in contrast to most existing methods that focus on the covariance matrix parameter requiring pre-set values for the others. An illustration on synthetic and real data is provided in the case of the t-distribution with unknown mean and degrees-of-freedom parameters.

6.3.2 Bayesian nonparametric models for hidden Markov random fields on non-Gaussian variables and applications

Participants Julyan Arbel, Jean-Baptiste Durand, Florence Forbes.

Joint work with: Hien Nguyen from La Trobe University Melbourne Australia and Grégoire Vincent from IRD, AMAP, Montpellier, France

Hidden Markov random fields (HMRFs) have been widely used in image segmentation and more generally, for clustering of data indexed by graphs. Dependent hidden variables (states) represent the cluster identities and determine their interpretations. Dependencies between state variables are induced by the notion of neighborhood in the graph. A difficult and crucial problem in HMRFs is the identification of the number of possible states K. Recently, selection methods based on Bayesian non parametric priors (Dirichlet processes) have been developed. They do not assume that K is bounded a priori, thus allowing its adaptive selection with respect to the quantity of available data and avoiding costly systematic estimation and comparison of models with different fixed values for K. Our previous work [27] has focused on Bayesian nonparametric priors for HMRFs and continuous, Gaussian observations. In this work, we consider extensions to non-Gaussian observed data. A first case is discrete data, typically issued from counts. A second is exponential-distributed data. We defined and implemented Bayesian nonparametric models for HMRFs with Poisson- and exponential-distributed observations. Inference is achieved by Variational Bayesian Expectation Maximization (VBEM).

We proposed an application of the discrete-data model to a new risk mapping model for traffic accidents in the region of Victoria, Australia [54]. The partition into regions using labels yielded by HMRFs was interpreted using covariates, which showed a good discrimination with regard to labels.

As a perspective, Bayesian nonparametric models for hidden Markov random fields could be extended to non-Poissonian models (particularly to account for zero-inflated and over-/under-dispersed cases of application) and to regression models.

The exponential model was applied to leaf density estimation in forests and isolated trees subjected to laser scans. The data are lengths of portions of laser beams between two hits of translucent materials (mainly, leaves). The sampling space is discretized into voxels and under some specific assumptions, the lengths have an exponential distribution with possible censoring if the beam leaves the voxel. The added-value of HMRFs is to go beyond the assumption of independent voxels and taking into account spatial dependencies between them, which are due to the underlying geometric structure of trees.

Current perspectives of this work include the improvement of the convergence in the VBEM algorithm, since however the KL divergence between the posterior distribution and its approximation converges, the sequence of optimizing parameters is shown to diverge in our current approach.

6.3.3 Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia

Participants Jean-Baptiste Durand, Florence Forbes.

Joint work with: Hien Nguyen, Long Truong, Q. Phan from La Trobe University Melbourne Australia.

We investigate the use of Bayesian nonparametric (BNP) models coupled with Markov random fields (MRF) in a risk mapping context, to build partitions of the risk into homogeneous spatial regions. In contrast to most existing methods, the proposed approach does not require an arbitrary commitment to a specified number of risk classes and determines their risk levels automatically. We consider settings in which the relevant information are counts and propose a so called BNP Hidden MRF (BNP-HMRF) model that is able to handle such data. The model inference is carried out using a variational Bayes Expectation–Maximisation algorithm and the approach is illustrated on traffic crash data in the state of Victoria, Australia. The obtained results corroborate well with the traffic safety literature. More generally, the model presented here for risk mapping offers an effective, convenient and fast way to conduct partition of spatially localised count data. Details can be found in [54].

6.3.4 Hidden Markov models for the analysis of eye movements

Participants Jean-Baptiste Durand, Sophie Achard.

Joint work with: Anne Guérin-Dugué (GIPSA-lab) and Benoit Lemaire (Laboratoire de Psychologie et Neurocognition)

This research theme is supported by a LabEx PERSYVAL-Lab project-team grant.

In the last years, GIPSA-lab has developed computational models of information search in web-like materials, using data from both eye-tracking and electroencephalograms (EEGs). These data were obtained from experiments, in which subjects had to decide whether a text was related or not to a target topic presented to them beforehand. In such tasks, reading process and decision making are closely related. Statistical analysis of such data aims at deciphering underlying dependency structures in these processes. Hidden Markov models (HMMs) have been used on eye-movement series to infer phases in the reading process that can be interpreted as strategies or steps in the cognitive processes leading to decision. In HMMs, each phase is associated with a state of the Markov chain. The states are observed indirectly though eye-movements. Our approach was inspired by Simola *et al.* (2008) [70], but we used hidden semi-Markov models for better characterization of phase length distributions (Olivier *et al.*, 2017) [69]. The estimated HMM highlighted contrasted reading strategies, with both individual and document-related variability. New results were obtained in the standalone analysis of the eye-movements: 1) a statistical comparison between the effects of three types of texts was performed, considering texts either closely related, moderately related or unrelated to the target topic; 2) a characterization of the effects of the distance to trigger words on transition probabilities and 3) highlighting a predominant intra-individual variability in scanpaths.

Our goal for this coming year is to use the segmentation induced by our eye-movement model to obtain a statistical characterization of functional brain connectivity through simultaneous EEG recordings. This should lead to some integrated models coupling EEG and eye movements within one single HMM for better identification of strategies.

6.3.5 Assessing spatial dependencies in the effect of treatment on neurite growth

Participants Jean-Baptiste Durand, Sophie Achard, Jiajie Li.

Joint work with: Stéphane Belin, Homaira Nawabi, Sabine Chierici from Grenoble Institute of Neuroscience.

The World Health Organization estimates that 250 000 to 500 000 new cases of spinal cord injuries occur each year. People suffering from those lesions endure irreversible disabilities, as no treatment is available to counteract the regenerative failure of mature Central Nervous System (CNS). Thus, promoting neuronal growth, repair and functional recovery remains one of the greatest challenges for neurology, patients and public health. Our partners at GIN (Grenoble Institute for Neurosciences) demonstrated that doublecortin is a key factor for axon regeneration and neuronal survival. Short peptides could be used as a treatment to enhance axon outgrowth. To test their potential effect on axonal growth, embryonic neurons in culture are treated with those peptides. Neurons are then imaged and neurite length is quantified automatically. The analysis of such data raises statistical questions to avoid bias in testing the relevance of a given peptide. All neuronal cultures are not the same. Particularly, the proximity between neurons is variable and likely to influence its intrinsic capability to grow. In such contexts, the usual test-based methodology to compare treatments cannot be applied and has to be adapted.

In this work, we highlighted the first-order spatial stationarity of neurite lengths within a same experiment, using HMRF models depicted in Subsection 6.3.3. Then we investigate spatial dependencies between lengths of close neurites, highlighting the relevance of CAR models of J. Besag to account for the effect of neighbours' lengths. This raises the question of choosing a relevant graph of dependencies in CAR and several types of graphs were compared.

6.3.6 Modelling the effects of cultivars and treatments on the structure of apple trees

Participants Jean-Baptiste Durand.

Joint work with: Evelyne Costes, INRAE, AGAP, Montpellier, France and Martin Mészáros, Research and Breeding Institute of Pomology Holovousy Ltd., Hořice, Czech Republic.

This study aims at characterizing the effects of cultivars and treatments on the structure of apple trees. More specifically, tree trunks are issued from the following cultivars Rubinola, Topaz and Golden Delicious. Each tree is fertilized one among these different nitrogen (N) doses: I) untreated (control), II) treated with 20 g N/tree/year, and III) treated with 30 g N/tree/year. We developed a modelling strategy inspired by Meszaros in 2020: it is assumed that in every cultivar under every treatment, there exists an underlying common sequence of development, which is indirectly observed through the following features attached to each metamer (elementary entity) of the trunk: length class of axillary shoots, together with their lateral and terminal flowering. This sequence is modelled by successions of zones along the trunks (zone lengths, transitions and distributions of features within each zone). These assumptions lead to estimated hidden semi-Markov chain (HSMC) models with similar definitions as in Subsection 6.3.4.

It now remains to model how the HSMC parameters depend on cultivars and treatments, which is planned to be handled with generalized linear models.

6.3.7 Splitting models for multivariate count data

Participants Jean-Baptiste Durand.

Joint work with: Jean Peyhardi, Institut Montpelliérain Alexander Grothendieck, Montpellier, France and Pierre Fernique, CIRAD, Agap, Montpellier, France

Modelling multivariate count data and their dependencies is a difficult problem, in absence of a reference probabilistic model equivalent to the multivariate Gaussian in the continuous case, which allows modelling arbitrary marginal and conditional independence properties among those representable by graphical models while keeping probabilistic computations tractable (or even better, explicit).

In this work, we investigated the class of splitting distributions as the composition of a singular multivariate distribution and a univariate distribution. It was shown that most common parametric count distributions (multinomial, negative multinomial, multivariate hypergeometric, multivariate negative hypergeometric, ...) can be written as splitting distributions with separate parameters for both components, thus facilitating their interpretation, inference, the study of their probabilistic characteristics and their extensions to regression models. We highlighted many probabilistic properties deriving from the compound aspect of splitting distributions and their underlying algebraic properties. Parameter inference and model selection are thus reduced to two separate problems, preserving time and space complexity of the base models. Based on this principle, we introduced several new distributions. In the case of multinomial splitting distributions, conditional independence and asymptotic normality properties for estimators were obtained. Mixtures of splitting regression models were used on a mango tree dataset in order to analyse its patchiness.

Conditional independence properties of estimators were obtained for sum and singular distribution parameters for MLE and Bayesian estimators in the framework of multinomial splitting distributions [29]. As a perspective, similar properties remain to be investigated for other cases of splitting (or possibly sum) distributions and regression models. Moreover, this work could be used for learning graphical models with discrete variables, which is an open issue. Although the graphical models for usual additive convolution splitting distributions are trivial (either complete or empty), they could be used as building blocks for partially directed acyclic graphical models. Therefore, some existing procedures for learning partially directed acyclic graphical models could be used for learning those based on convolution splitting distributions and regressions. Such approaches could be used for instance to infer gene co-expression network from RNA seq data sets.

6.3.8 Bayesian neural networks

Participants Julyan Arbel, Mariia Vladimirova.

Joint work with: Pablo Mesejo from University of Granada, Spain, Jakob Verbeek from Inria Grenoble Rhône-Alpes, France.

We investigate deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities, shedding light on novel sparsity-inducing mechanisms at the level of the units of the network, both pre- and post-nonlinearities. The main thrust of the paper is to establish that the units prior distribution becomes increasingly heavy-tailed with depth. We show that first layer units are Gaussian, second layer units are sub-Exponential, and we introduce sub-Weibull distributions to characterize the deeper layers units. Bayesian neural networks with Gaussian priors are well known to induce the weight decay penalty on the weights. In contrast, our result indicates a more elaborate regularisation scheme at the level of the units. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.

6.3.9 Brain connectivity

Participants Sophie Achard.

Joint work with: Emmanuel Barbier from GIN and Guillaume Becq from GIPSA-lab, Univ. Grenoble Alpes

In two recent publications [16] and [16], we evaluated the reliability of graph connectivity estimations using wavelets. Under anesthesia, systemic variables and CBF are modified. How does this alter the connectivity measures obtained with rs-fMRI? To tackle this question, we explored the effect of four different anesthetics on Long Evans and Wistar rats with multimodal recordings of rs-fMRI, systemic variables and CBF. After multimodal signal processing, we show that the blood-oxygen-level-dependent (BOLD) variations and functional connectivity (FC) evaluated at low frequencies (0.031–0.25 Hz) do not depend on systemic variables and are preserved across large interval of baseline CBF values. Based on these findings, we found that most brain areas remain functionally active under any anesthetics, i.e. connected to at least one other brain area, as shown by the connectivity graphs. In addition, we quantified the influence of nodes by a measure of functional connectivity strength to show the specific areas targeted by anesthetics and compare correlation values of edges at different levels. These measures enable us to highlight the specific network alterations induced by anesthetics. Altogether, this suggests that changes in connectivity could be evaluated under anesthesia, routinely used in the control of neurological injury.

7 Bilateral contracts and grants with industry

7.1 Bilateral contracts with industry

Contract with EDF (2020-2023). Julyan Arbel and Stéphane Girard are the advisors of the PhD thesis of Théo Moins founded by EDF. The goal is to investigate sensitivity analysis and extrapolation limits in extreme-value theory Bayesian methods. The financial support for STATIFY is of 150 keuros.

Contract with TDK-Invensense (2020-2023). Julyan Arbel is the advisor of the PhD thesis of Minh Tri Lê founded by TDK-Invensense. The goal is to apply deep learning methods on small size systems, thus investigating compression methods in deep learning. The financial support for STATIFY is of 150 keuros.

Contract with VALEO (2018-2020). Stéphane Girard and Pascal Dkengne Sielenou are involved in a study with Valeo to assess the relevance of extreme-value theory in the calibration of sensors for autonomous cars. The financial support for STATIFY is of 200 keuros.

8 Partnerships and cooperations

8.1 International initiatives

8.1.1 Inria International Labs

STATIFY is involved in the Inria associate team SIMERG2E (Statistical Inference for the Management of Extreme Risks, Genetics and Global Epidemiology) headed by Stéphane Girard, 2015-2020, part of the LIRIMA international lab, and together with LERSTAD, Université Gaston Berger (Senegal). Two research axes are explored: 1) Spatial extremes, application to management of extreme risks. We address the definition of new risk measures, the study of their properties in case of extreme events and their estimation from data and covariate information. Our goal is to obtain estimators accounting for possible variability, both in terms of space and time, which is of prime importance in many hydrological, agricultural and energy contexts. 2) Classification, application to genetics and global epidemiology. We address the challenge to build statistical models in order to test association between diseases and human host genetics in a context of genome-wide screening. Adequate models should allow to handle complexity in genomic data (correlation between genetic markers, high dimensionality) and additional statistical issues present in data collected from a family-based longitudinal survey (non-independence between individuals due to familial relationship and non-independence within individuals due to repeated measurements on a same person over time).

8.1.2 Inria associate team not involved in an IIL

LANDER: Title: Latent Analysis, Adversarial Networks, and DimEnsionality Reduction - International Partner (Institution - Laboratory - Researcher): Start year: 2019. See also: https://team.inria.fr/mistis/projects/lander/

The collaboration is based on three main points, in statistics, machine learning and applications: 1) clustering and classification (mixture models), 2) regression and dimensionality reduction (mixture of regression models and non parametric techniques) and 3) high impact applications (neuroimaging and MRI). Our overall goal is to collectively combine our resources and data in order to develop tools that are more ubiquitous and universal than we could have previously produced, each on our own. A wide class of problems from medical imaging can be formulated as inverse problems. Solving an inverse problem means recovering an object from indirect noisy observations. Inverse problems are therefore often compounded by the presence of errors (noise) in the data but also by other complexity sources such as the high dimensionality of the observations and objects to recover, their complex dependence structure and the issue of possibly missing data. Another challenge is to design numerical implementations that are computationally efficient. Among probabilistic models, generative models have appealing properties to meet all the above constraints. They have been studied in various forms and rather independently both in the statistical and machine learning literature with different depths and insights, from the well established probabilistic graphical models to the more recent (deep) generative adversarial networks (GAN). The advantages of the latter being primarily computational and their disadvantages being the lack of theoretical statements, in contrast to the former. The overall goal of the collaboration is to build connections between statistical and machine learning tools used to construct and estimate generative models with the resolution of real life inverse problems as a target. This induces in particular the need to help the models scale to high dimensional data while maintaining our ability to assess their correctness, typically the uncertainty associated to the provided solutions.

Informal international partners The context of our research is also the collaboration between STATIFY and a number of international partners.

The main other active international collaborations in 2020 are with:

- E. Deme and A. Diop from Gaston Berger University in Senegal.
- Guillaume Kon Kam King, Stefano Favaro, Pierpaolo De Blasi, Collegio Carlo Alberto, Turin, Italy.
- Igor Prünster, Antonio Lijoi, and Riccardo Corradin Bocconi University, Milan, Italy.
- Bernardo Nipoti, Trinity College Dublin, Ireland.
- Stephen Walker, University of Texas at Austin, USA.
- Alex Petersen, University of California Santa Barbara, USA.
- Dimitri van de Ville, EPFL, University of Geneva, Switzerland.

8.1.3 Participation in other international programs

Sophie Achard is coPI of the ANR project (PRCI) QFunC in partnership with University of Santa Barbara (USA) and Université de Lausanne (Switzerland). The aim of the project is to build spatio-temporal models for brain connectivity. The financial support for Statify is 260000 euros.

8.2 National initiatives

ANR STATIFY is involved in the 4-year ANR project ExtremReg (2019-2023) hosted by Toulouse University. This research project aims to provide new adapted tools for nonparametric and semiparametric modeling from the perspective of extreme values. Our research program concentrates around three central themes. First, we contribute to the expanding literature on non-regular boundary regression where smoothness and shape constraints are imposed on the regression function and the regression errors are not assumed to be centred, but one-sided. Our second aim is to further investigate the study of the modern extreme value theory built on the use of asymmetric least squares instead of traditional quantiles and order statistics. Finally, we explore the less-discussed problem of estimating high-dimensional, conditional and joint extremes

The financial support for STATIFY is about 15.000 euros.

STATIFY is also involved in the ANR project GAMBAS (2019-2023) hosted by Cirad, Montpellier. The project Generating Advances in Modeling Biodiversity And ecosystem Services (GAMBAS) develops statistical improvements and ecological relevance of joint species distribution models. The project supports the PhD thesis of Giovanni Poggiato.

Grenoble Idex projects STATIFY is involved in a transdisciplinary project **NeuroCoG** and in a newly accepted cross-disciplinary project (CDP) **Risk@UGA**. F. Forbes is also a member of the executive committee and responsible for the *Data Science for life sciences* work package in another project entitled **Grenoble Alpes Data Institute**.

- The main objective of the RISK@UGA project is to provide some innovative tools both for the management of risk and crises in areas that are made vulnerable because of strong interdependencies between human, natural or technological hazards, in synergy with the conclusions of Sendai conference. The project federates a hundred researchers from Human and Social Sciences, Information & System Sciences, Geosciences and Engineering Sciences, already strongly involved in the problems of risk assessment and management, in particular natural risks. The PhD thesis of Meryem Bousebata is one of the eleven PhDs funded by this project.
- The NeuroCoG project aims at understanding the biological, neurophysiological and functional bases of behavioral and cognitive processes in normal and pathological conditions, from cells to networks and from individual to social cognition. No decisive progress can be achieved in this area without an aspiring interdisciplinary approach. The interdisciplinary ambition of NeuroCoG is

particularly strong, bringing together the best scientists, engineers and clinicians at the crossroads of experimental and life sciences, human and social sciences and information and communication sciences, to answer major questions on the workings of the brain and of cognition. One of the work package entitled InnobioPark is dedicated to Parkinson's Disease. The PhD thesis of Veronica Munoz Ramirez is one of the three PhDs in this work package.

- The Grenoble Alpes Data Institute aims at undertaking groundbreaking interdisciplinary research focusing on how data change science and society. It combines three fields of data-related research in a unique way: data science applied to spatial and environmental sciences, biology, and health sciences; data-driven research as a major tool in Social Sciences and Humanities; and studies about data governance, security and the protection of data and privacy. In this context, a 2-year multi-disciplinary projects has been granted in November 2018 to Mistis in collaboration with the Grenoble Institute of Neuroscience. The objective of this project is to develop a statistical learning technique that is able to solve a problem of tracking and analyzing a large population of single molecules. The main difficulties are: 1) the large number of observations to analyse, 2) the noisy nature of the signals, 3) the definition of a quality index to allow the elimination of poor-quality data and false positive signals. We also aim at providing a powerful, well-documented and open-source software, that will be user-friendly for non-specialists.

In the context of the Idex associated with the Université Grenoble Alpes, Alexandre Constantin was awarded half a PhD funding from IRS (Initiatives de Recherche Stratégique), 50 keuros.
In the context of the MIAI (Multidisciplinary Institute in Artificial Intelligence) institute and its open call to sustain the development and promotion of AI, Stéphane Girard was awarded a grant of 4500 euros for his project "Simulation of extreme values by AI generative models. Application to banking risk" joint with CMAP, Ecole Polytechnique.
In the context of the MIAI (Multidisciplinary Institute in Artificial Intelligence) institute and its open call to sustain the development and promotion of AI, Julyan Arbel was awarded a grant of 5000 euros for his project "Bayesian deep learning".
Julyan Arbel was awarded a grant of 10000 euros for his project "Bayesian nonparametric modeling".

8.2.1 Networks

MSTGA and AIGM INRA (French National Institute for Agricultural Research) networks: F. Forbes and J.B Durand are members of the INRA network called AIGM (ex MSTGA) network since 2006, http://carlit.toulouse.inra.fr/AIGM, on Algorithmic issues for Inference in Graphical Models. It is funded by INRA MIA and RNSC/ISC Paris. This network gathers researchers from different disciplines. MISTIS co-organized and hosted 2 of the network meetings in 2008 and 2015 in Grenoble.

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

General chair, scientific chair

- Florence Forbes was a member of the scientific committee of Bayes Comp 2020, a biennial conference sponsored by the ISBA section of the same name link.
- Julyan Arbel was a member of the scientific committee of Statistical Methods for Post Genomic Data analysis (SMPGD), link, and a member of the scientific program committee of the International Conference of Computational and Methodological Statistics (CMStat) link.

Member of the organizing committees

- Julyan Arbel organized the session entitled 'Bayesian Machine Learning' at the 13th International Conference of Computational and Methodological Statistics (CMStat), University of London, UK (December 2020).

- Julyan Arbel is organizing bi-weekly One World seminar on Approximate Bayesian computation (ABC) link.

9.1.2 Journal

Member of the editorial boards

- Julyan Arbel is Associate Editor of *Bayesian Analysis* since 2019.
- Julyan Arbel and Florence Forbes are Associate Editors of *Australian and New Zealand Journal of Statistics* since 2019.
- Julyan Arbel is Associate Editor of *Statistics & Probability Letters* since 2019.
- Julyan Arbel is Associate Editor of *Computational Statistics & Data Analysis* since 2020.
- Florence Forbes is Associate Editor of *Computational Statistics & Data Analysis* since 2017.
- Stéphane Girard is Associate Editor of *Dependence Modelling* (De Gruyter) since 2015.
- Stéphane Girard is Associate Editor of *Journal of Multivariate Analysis* (Elsevier) depuis 2016.
- Stéphane Girard is Associate Editor of *Revstat - Statistical Journal* since 2019.

Reviewer - reviewing activities

- Julyan Arbel has been a rewiewer for the Annals of Statistics, Biometrika, JASA (Journal of the American Statistical Association), Journal of Multivariate Analysis, Scandinavian Journal of Statistics, Stochastic Processes and their Applications, IEEE Transactions on Signal Processing, IEEE Access, Econometrics and Statistics, and for the following machine learning conferences: Conference on Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML), Conference on Learning Representations (ICLR), Symposium on Advances in Approximate Bayesian Inference (AABI), and for the US Army Research Office (ARO), and for a book on Bayesian nonparametrics published by CRC Press.
- Stéphane Girard has been a rewiewer for JASA (Journal of the American Statistical Association), JSPI (Journal of Statistical Planning and Inference) and EJS (Electronic Journal of Statistics).
- Florence Forbes has been a reviewer for IEEE Transactions PAMI and Statistics & Computing.

9.1.3 Invited talks

- Julyan Arbel was an invited speaker at 13th International Conference of Computational and Methodological Statistics (CMStat), University of London, UK, December.
- Julyan Arbel was an invited speaker at Statistics Seminar Series in the School of Mathematics & Statistics, University College Dublin, November 23.
- Stéphane Girard was an invited speaker at StressTest-2020: International Workshop on Stress Test and Risk Management [42].
- Florence Forbes was an invited speaker at the AppliBUGS group Seminar (link).

9.1.4 Scientific expertise

- Julyan Arbel is a scientific committee member of the Data Science axis of Persyval Labex (Machine learning: fundamentals and applications, and Data linking, sharing and privacy), since 2019.
- Stéphane Girard was a member of the committee in charge of hiring a Professor at LJK, Université Grenoble-Alpes.
- Stéphane Girard acted as an expert of NWO projects evaluation (Netherlands Organisation for Scientific Research).
- Florence Forbes was a member of the committee in charge of hiring professors and teaching assistants at Ecole Polytechnique, Paris.
- Florence Forbes acted as an expert for the tenure application of Lo Bin Chang at Ohio University, USA.
- Florence Forbes acted as a reviewer for Charles University, Prague.
- Florence Forbes is a member of the Helmholtz AI Cooperation Unit advisory committee (link), 2019-present.
- Florence Forbes is a member of the EURASIP Technical Area Committee BISA (Biomedical Image & Signal Analytics) since January 2021 for a 3 years duration.

9.2 Teaching - Supervision - Juries

9.2.1 Teaching

- Licence : Stéphane Girard, *Principe et Méthodes Statistiques*, 18 ETD, L3 level, Ensimag. Grenoble-INP, France.
- Master : Stéphane Girard, *Statistique Inférentielle Avancée*, 18 ETD, M1 level, Ensimag. Grenoble-INP, France.
- Master and PhD course: Julyan Arbel, *Bayesian nonparametrics and Bayesian deep learning*, Master Mathématiques Apprentissage et Sciences Humaines (M*A*S*H), Université PSL (Paris Sciences & Lettres), 25 ETD. *Bayesian deep learning*, Master Intelligence Artificielle, Systèmes, Données (IASD), Université PSL (Paris Sciences & Lettres), 12 ETD.
- Master and PhD course: Julyan Arbel, *Bayesian machine learning*, Master Mathématiques Vision et Apprentissage Master MVA, École normale supérieure Paris-Saclay, 36 ETD.
- Master: Jean-Baptiste Durand, *Statistics and probability*, 192H, M1 and M2 levels, Ensimag Grenoble INP, France. Head of the MSIAM M2 program, in charge of the data science track.
- Jean-Baptiste Durand is a faculty member at Ensimag, Grenoble INP.
- Sophie Achard M1 course Théorie des graphes et réseaux sociaux, M1 level, MIASHS, Université Grenoble Alpes (UGA), 14 ETD.

9.2.2 Supervision

- PhD defended: Aboubacrène Ag Ahmad "*Modélisation semi-paramétrique des extrêmes conditionnels*", [46], September 2020, Stéphane Girard and Alio Diop, Université Gaston Berger, Sénégal.
- PhD defended: Veronica Munoz,"*Extraction de signatures dans les données IRM de patients parkinsoniens de novo*", Florence Forbes and Michel Dojat, Université Grenoble Alpes, December 2020.
- PhD defended: Fabien Boux,"*Développement de méthodes statistiques pour l'imagerie IRM fingerprinting*", Florence Forbes and Emmanuel Barbier, Université Grenoble Alpes, December 2020.

- PhD in progress: Benoit Kugler, "*Massive hyperspectral images analysis by inverse regression of physical models*", Florence Forbes and Sylvain Douté, Université Grenoble Alpes, started on October 2018.
- PhD in progress: Mariia Vladimirova, "*Prior specification for Bayesian deep learning models and regularization implications*", started on October 2018, Julyan Arbel and Jakob Verbeek, Université Grenoble Alpes.
- PhD in progress: Théo Moins "*Quantification bayésienne des limites d'extrapolation en statistique des valeurs extrêmes*", started on October 2020, Stéphane Girard and Julyan Arbel, Université Grenoble Alpes.
- PhD in progress: Michael Allouche "*Simulation d'extrêmes par modèles génératifs et applications aux risques bancaires*", started on April 2020, Stéphane Girard and Emmanuel Gobet, Ecole Polytechnique.
- PhD in progress: Meryem Bousebata "*Bayesian estimation of extreme risk measures: Implication for the insurance of natural disasters*", started on October 2018, Stéphane Girard and Geffroy Enjolras, Université Grenoble Alpes.
- PhD in progress: Alexandre Constantin "*Analyse de séries temporelles massives d'images satellitaires : Applications à la cartographie des écosystèmes*", started on November 2018, Stéphane Girard and Mathieu Fauvel, Université Grenoble Alpes.
- PhD in progress: Daria Bystrova, "*Joint Species Distribution Modeling: Dimension reduction using Bayesian nonparametric priors*", started on October 2019, Julyan Arbel and Wilfried Thuiller, Université Grenoble Alpes.
- PhD in progress: Giovanni Poggiatto, "*Scalable Approaches for Joint Species Distribution Modeling*", started on November 2019, Julyan Arbel and Wilfried Thuiller, Université Grenoble Alpes.
- PhD in progress: Minh Tri Lê, "*Constrained signal processing using deep neural networks for MEMs sensors based applications.*", started on September 2020, Julyan Arbel and Etienne de Foras, Université Grenoble Alpes, CIFRE Invensense.

9.2.3 Juries

- Florence Forbes has been reviewer for the PhD thesis of Faustine Bousquet (Universite de Montpellier), Raphaelle Momal (Université Paris-Saclay), Nikola Hrelja (Ecole Polytechnique), Maxime Vono (Université de Toulouse) and Thi Khuyen Le (Université Aix-Marseille).
- Florence Forbes has been a member of the PhD committee of Bruno Meriaux (Université Paris-Saclay) and of the HDR committee of Jean-Baptiste Durand (Université Grenoble Alpes).
- Sophie Achard has been reviewer for the HDR of Julien Modolo (Université Rennes 1).

10 Scientific production

10.1 Major publications

[1] C. Amblard and S. Girard. 'Estimation procedures for a semiparametric family of bivariate copulas'. In: *Journal of Computational and Graphical Statistics* 14.2 (2005), pp. 1–15.

[2] J. Blanchet and F. Forbes. 'Triplet Markov fields for the supervised classification of complex structure data'. In: *IEEE trans. on Pattern Analyis and Machine Intelligence* 30(6) (2008), pp. 1055–1067.

[3] C. Bouveyron, S. Girard and C. Schmid. 'High dimensional data clustering'. In: *Computational Statistics and Data Analysis* 52 (2007), pp. 502–519.

[4] C. Bouveyron, S. Girard and C. Schmid. 'High dimensional discriminant analysis'. In: *Communication in Statistics - Theory and Methods* 36.14 (2007).

[5] L. Chaari, T. Vincent, F. Forbes, M. Dojat and P. Ciuciu. 'Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach'. In: *IEEE Transactions on Medical Imaging* 32.5 (May 2013), pp. 821–837. DOI: 10.1109/TMI.2012.2225636. URL: http://hal.inria.fr/inserm-00753873.

[6] A. Daouia, S. Girard and G. Stupfler. 'Estimation of Tail Risk based on Extreme Expectiles'. In: *Journal of the Royal Statistical Society series B* 80 (2018), pp. 263–292.

[7] A. Deleforge, F. Forbes and R. Horaud. 'High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables'. In: *Statistics and Computing* (Feb. 2014). DOI: 10.1007/s11222-014-9461-5. URL: https://hal.inria.fr/hal-00863468.

[8] F. Forbes and G. Fort. 'Combining Monte Carlo and Mean field like methods for inference in hidden Markov Random Fields'. In: *IEEE trans. Image Processing* 16.3 (2007), pp. 824–837.

[9] F. Forbes and D. Wraith. 'A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering'. In: *Statistics and Computing* 24.6 (Nov. 2014), pp. 971–984. DOI: 10.1007/s11222-013-9414-4. URL: https://hal.inria.fr/hal-00823451.

[10] S. Girard. 'A Hill type estimate of the Weibull tail-coefficient'. In: *Communication in Statistics - Theory and Methods* 33.2 (2004), pp. 205–234.

10.2 Publications of the year

International journals

[11] A. Ag Ahmad, H. Deme, A. Diop, S. Girard and A. Usseglio-Carleve. 'Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model'. In: *Electronic journal of statistics* 14.2 (2020), pp. 4421–4456. DOI: 10.1214/20-EJS1779. URL: https://hal.inria.fr/hal-02486937.

[12] C. Albert, A. Dutfoy, L. Gardes and S. Girard. 'An extreme quantile estimator for the log-generalized Weibull-tail model'. In: *Econometrics and Statistics* 13 (Jan. 2020), pp. 137–174. DOI: 10.1016/j.ecosta.2019.01.004. URL: https://hal.inria.fr/hal-01783929.

[13] C. Albert, A. Dutfoy and S. Girard. 'Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles'. In: *Extremes* 23 (June 2020), pp. 349–380. DOI: 10.1007/s10687-019-00370-2. URL: https://hal.archives-ouvertes.fr/hal-01692544.

[14] J. Arbel, R. Corradin and B. Nipoti. 'Dirichlet process mixtures under affine transformations of the data'. In: *Computational Statistics* 36 (Mar. 2021), pp. 577–601. DOI: 10.1007/s00180-020-01013-y. URL: https://hal.archives-ouvertes.fr/hal-01950652.

[15] J. Arbel, O. Marchal and B. Nipoti. 'On the Hurwitz zeta function with an application to the exponential-beta distribution'. In: *Journal of Inequalities and Applications* (11th Feb. 2020). DOI: 10.1186/s13660-020-02357-1. URL: https://hal.archives-ouvertes.fr/hal-02400451.

[16] G. J.-P. C. Becq, E. L. Barbier and S. Achard. 'Brain networks of rats under anesthesia using resting-state fMRI: comparison with dead rats, random noise and generative models of networks'. In: *Journal of Neural Engineering* 17.4 (1st Aug. 2020), p. 045012. DOI: 10.1088/1741-2552/ab9fec. URL: https://hal.archives-ouvertes.fr/hal-02935391.

[17] G. J.-P. C. Becq, T. Habet, N. Collomb, M. Faucher, C. Delon-Martin, V. Coizet, S. Achard and E. L. Barbier. 'Functional connectivity is preserved but reorganized across several anesthetic regimes'. In: *NeuroImage* 219 (Oct. 2020), p. 116945. DOI: 10.1016/j.neuroimage.2020.116945. URL: https://hal.archives-ouvertes.fr/hal-02935430.

[18] F. Boux, F. Forbes, J. Arbel, B. Lemasson and E. L. Barbier. 'Bayesian inverse regression for vascular magnetic resonance fingerprinting'. In: *IEEE Transactions on Medical Imaging* (2021). URL: https://hal.archives-ouvertes.fr/hal-02314026.

[19] F. Boux, F. Forbes, N. Collomb, e. zub emma, L. Maziere, F. Bock, M. Blaquière, V. Stupar, A. Depaulis, N. Marchi and E. Barbier. 'Neurovascular multiparametric MRI defines epileptogenic and seizure propagation regions in experimental mesiotemporal lobe epilepsy'. In: *Epilepsia* (Apr. 2021). DOI: 10.1111/epi.16886. URL: https://hal.archives-ouvertes.fr/hal-03193021.

[20] C. Brossard, O. Montigon, F. Boux, A. Delphin, T. Christen, E. L. Barbier and B. Lemasson. 'MP3: Medical software for Processing multi-Parametric images Pipelines'. In: *Frontiers in Neuroinformatics* 14 (16th Nov. 2020). DOI: 10.3389/fninf.2020.594799. URL: https://hal.archives-ouvertes.fr/hal-02937866.

[21] D. Bystrova, G. Poggiato, B. Bektaş, J. Arbel, J. S. Clark, A. Guglielmi and W. Thuiller. 'Clustering species with residual covariance matrix in Joint Species Distribution models'. In: *Frontiers in Ecology and Evolution* (2021), pp. 1–20. DOI: 10.3389/fevo.2021.601384. URL: https://hal.inria.fr/hal-03151472.

[22] A. Daouia, S. Girard and G. Stupfler. 'ExpectHill estimation, extreme risk and heavy tails'. In: *Journal of Econometrics* (2020), pp. 1–54. DOI: 10.1016/j.jeconom.2020.02.003. URL: https://hal.inria.fr/hal-01856212.

[23] A. Daouia, S. Girard and G. Stupfler. 'Tail expectile process and risk assessment'. In: *Bernoulli* 26.1 (2020), pp. 531–556. DOI: 10.3150/19-BEJ1137. URL: https://hal.archives-ouvertes.fr/hal-01744505.

[24] L. Gardes and S. Girard. 'On the estimation of the variability in the distribution tail'. In: *Test* (2021). DOI: 10.1007/s11749-021-00754-2. URL: https://hal.inria.fr/hal-02400320.

[25] L. Gardes, S. Girard and G. Stupfler. 'Beyond tail median and conditional tail expectation: extreme risk estimation using tail L^p–optimisation'. In: *Scandinavian Journal of Statistics* 47.3 (Sept. 2020), pp. 922–949. DOI: 10.1111/sjos.12433. URL: https://hal.inria.fr/hal-01726328.

[26] S. Girard, G. Stupfler and A. Usseglio-Carleve. 'Nonparametric extreme conditional expectile estimation'. In: *Scandinavian Journal of Statistics* (2020). DOI: 10.1111/sjos.12502. URL: https://hal.archives-ouvertes.fr/hal-02114255.

[27] H. Lu, J. Arbel and F. Forbes. 'Bayesian nonparametric priors for hidden Markov random fields'. In: *Statistics and Computing* 30 (2020), pp. 1015–1035. DOI: 10.1007/s11222-020-09935-9. URL: https://hal.archives-ouvertes.fr/hal-02163046.

[28] H. D. Nguyen, F. Forbes and G. J. Mclachlan. 'Mini-batch learning of exponential family finite mixture models'. In: *Statistics and Computing* 30 (July 2020), pp. 731–748. DOI: 10.1007/s11222-019-09919-4. URL: https://hal.archives-ouvertes.fr/hal-02415068.

[29] J. Peyhardi, P. Fernique and J.-B. Durand. 'Splitting models for multivariate count data'. In: *Journal of Multivariate Analysis* 181 (Jan. 2021), p. 104677. DOI: 10.1016/j.jmva.2020.104677. URL: https://hal.inria.fr/hal-02962877.

[30] G. Poggiato, T. Münkemüller, D. Bystrova, J. Arbel, J. Clark and W. Thuiller. 'On the interpretations of joint modelling in community ecology'. In: *Trends in Ecology and Evolution* (2021). DOI: 10.1016/j.tree.2021.01.002. URL: https://hal.archives-ouvertes.fr/hal-03153558.

[31] F. Privé, J. Arbel and B. J. Vilhjálmsson. 'LDpred2: better, faster, stronger'. In: *Bioinformatics* (16th Dec. 2020). DOI: 10.1093/bioinformatics/btaa1029. URL: https://hal.archives-ouvertes.fr/hal-03132949.

[32] F. Renard, C. Heinrich, M. Bouthillon, M. Schenck, F. Schneider, S. Kremer and S. Achard. 'A covariate-constraint method to map brain feature space into lower dimensional manifolds'. In: *Network Neuroscience* 5.1 (Jan. 2021), pp. 252–273. DOI: 10.1162/netn_a_00176. URL: https://hal.archives-ouvertes.fr/hal-03165916.

[33] M. Stehlik, J. Kiseľák, M. Vaičiulis, P. K. Jordanova, L. N. Soza, Z. Fabián, P. Hermann, L. Střelec, A. Rivera, S. Girard and S. Torres. 'Priority statement and some properties of t-lgHill estimator'. In: *Extremes* 23.3 (Sept. 2020), pp. 393–399. DOI: 10.1007/s10687-020-00375-2. URL: https://hal.inria.fr/hal-02540248.

[34] M. Vladimirova, S. Girard, H. Nguyen and J. Arbel. ‘Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions’. In: *Stat* (1st Oct. 2020). DOI: 10.1002/sta4.318. URL: https://hal.inria.fr/hal-02545121.

[35] F. Zheng, M. Jalbert, F. Forbes, S. Bonnet, A. Wojtusciszyn, S. Lablanche and P.-Y. Benhamou. ‘Characterization of Daily Glycemic Variability in Subjects with Type 1 Diabetes Using a Mixture of Metrics’. In: *Diabetes Technology and Therapeutics* 22.4 (Apr. 2020), pp. 301–313. DOI: 10.1089/dia.2019.0250. URL: https://hal.archives-ouvertes.fr/hal-02415078.

International peer-reviewed conferences

[36] M. Bousebata, G. Enjolras and S. Girard. ‘The dependence structure between yields and prices: A copula-based model of French farm income’. In: AAEA 2020 - Annual Meeting of the Agricultural and Applied Economics Association. Virtuel, United States, 10th Aug. 2020, pp. 1–15. DOI: 10.22004/ag.econ.304313. URL: https://hal.inria.fr/hal-02933766.

[37] D. Bystrova, J. Arbel, G. Kon Kam King and F. DESLANDES. ‘Approximating the clusters’ prior distribution in Bayesian nonparametric models’. In: AABI 2020 - 3rd Symposium on Advances in Approximate Bayesian Inference. Online, United States, 13th Jan. 2021, pp. 1–16. URL: https://hal.inria.fr/hal-03151483.

[38] G. Cathelain, B. Rivet, S. Achard, J. Bergounioux and F. Jouen. ‘Smart ballistocardiography front-end’. In: I2MTC 2020 - IEEE International Instrumentation and Measurement Technology Conference (I2MTC). Dubrovnik (online), Croatia: IEEE, 25th May 2020, pp. 1–6. DOI: 10.1109/I2MTC43012.2020.9128774. URL: https://hal.archives-ouvertes.fr/hal-03078465.

[39] G. Cathelain, B. Rivet, S. Achard, J. Bergounioux and F. Jouen. ‘U-Net Neural Network for Heartbeat Detection in Ballistocardiography’. In: EMBC 2020 - CMBEC 2020 - 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society - 43rd Annual Conference of the Canadian Medical and Biological Engineering Society. Montreal (virtual), Canada: IEEE, 20th July 2020, pp. 465–468. DOI: 10.1109/EMBC44109.2020.9176687. URL: https://hal.archives-ouvertes.fr/hal-03078490.

Conferences without proceedings

[40] A. A. Ahmad, S. Girard, A. Usseglio-Carleve, A. Diop and E. H. Deme. ‘Estimation of extreme quantiles from heavy-tailed distributions in a semi-parametric location-dispersion regression model’. In: Quatrièmes rencontres des jeunes chercheurs africains en France. Virtual, France, 10th Dec. 2020. URL: https://hal.inria.fr/hal-03065938.

[41] F. Boux, F. Forbes, J. Arbel, A. Delphin, T. Christen and E. Barbier. ‘Dictionary-based Learning in MR Fingerprinting: Statistical Learning versus Deep Learning’. In: International Society for Magnetic Resonance in Medicine (ISMRM). Sidney, Australia, 8th Aug. 2020. URL: https://hal.archives-ouvertes.fr/hal-02922858.

[42] S. Girard, A. A. Ahmad, E. H. Deme, A. Diop and A. Usseglio-Carleve. ‘Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model’. In: StressTest-2020 - International Workshop on Stress Test and Risk Management. Paris / Virtual, France, 30th Nov. 2020. URL: https://hal.inria.fr/hal-03040245.

[43] B. Kugler, F. Forbes and S. Douté. ‘An efficient Bayesian method for inverting physical models on massive planetary data’. In: EPSC 2020 - Europlanet Science Congress. Granada (Virtual meeting), Spain, 21st Sept. 2020, pp. 1–4. URL: https://hal.archives-ouvertes.fr/hal-02951518.

[44] B. Kugler, F. Forbes and S. Douté. ‘First order Sobol indices for physical models via inverse regression’. In: JDS 2020 - 52èmes Journées de Statistique de la Société Française de Statistique (SFdS). Nice, France, 7th June 2021, pp. 1–6. URL: https://hal.archives-ouvertes.fr/hal-02951375.

[45] A. Usseglio-Carleve, S. Girard and G. Stupfler. 'On automatic bias reduction for extreme expectile estimation'. In: CMStatistics 2020 - 13th International Conference of the ERCIM WG on Computational and Methodological Statistics. London / Virtual, United Kingdom, 19th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03087164.

Doctoral dissertations and habilitation theses

[46] A. A. Ahmad. 'Semi-parametric modelling of conditional extremes'. Université Gaston Berger de Saint-Louis (Sénégal), 16th Sept. 2020. URL: https://hal.inria.fr/tel-02941714.

Reports & preprints

[47] S. Achard, P. Borgnat and I. Gannaz. *Asymptotic control of FWER under Gaussian assumption: application to correlation tests.* 1st July 2020. URL: https://hal.archives-ouvertes.fr/hal-02883720.

[48] K. Ashurbekova, A. Usseglio-Carleve, F. Forbes and S. Achard. *Optimal shrinkage for robust covariance matrix estimators in a small sample size setting.* 27th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-02378034.

[49] M. Bousebata, G. Enjolras and S. Girard. *Extreme Partial Least-Squares regression.* 2021. URL: https://hal.inria.fr/hal-03165399.

[50] D. Bystrova, J. Arbel and T. Rahier. *Contributed comment on Article by Hahn, Murray, and Carvalho.* 12th Aug. 2020. URL: https://hal.archives-ouvertes.fr/hal-03149459.

[51] D. Bystrova, G. Poggiato, J. Arbel and W. Thuiller. *Latent factor models: a tool for dimension reduction in joint species distribution models.* 23rd Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03149452.

[52] A. Constantin, M. Fauvel and S. Girard. *Joint Supervised Classification and Reconstruction of Irregularly Sampled Satellite Image Times Series.* 2020. URL: https://hal.inria.fr/hal-02997573.

[53] P. S. Dkengne, S. Girard and S. Ahiad. *An automatic procedure to select a block size in the continuous generalized extreme value model estimation.* 29th Sept. 2020. URL: https://hal.inria.fr/hal-02952279.

[54] J.-B. Durand, F. Forbes, C. D. Phan, L. Truong, H. D. Nguyen and F. Dama. *Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia.* 4th Feb. 2021. URL: https://hal.inria.fr/hal-03138803.

[55] F. Forbes, H. D. Nguyen, T. T. Nguyen and J. Arbel. *Approximate Bayesian computation with surrogate posteriors.* 11th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03139256.

[56] S. Girard, G. Stupfler and A. Usseglio-Carleve. *An Lp –quantile methodology for estimating extreme expectiles.* 6th Apr. 2020. URL: https://hal.inria.fr/hal-02311609.

[57] S. Girard, G. Stupfler and A. Usseglio-Carleve. *Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models.* 22nd Apr. 2021. URL: https://hal.inria.fr/hal-02531027.

[58] S. Girard, G. Stupfler and A. Usseglio-Carleve. *Functional estimation of extreme conditional expectiles.* 1st Mar. 2021. URL: https://hal.inria.fr/hal-03117547.

[59] S. Girard, G. Stupfler and A. Usseglio-Carleve. *On automatic bias reduction for extreme expectile estimation.* 5th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03086048.

[60] B. Kugler, F. Forbes and S. Douté. *Fast Bayesian Inversion for high dimensional inverse problems.* 28th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02908364.

[61] V. Miele, C. Matias, M. Ohlmann, G. Poggiato, S. Dray and W. Thuiller. *Quantifying the overall effect of biotic interactions on species communities along environmental gradients.* 18th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03172480.

[62] H. D. Nguyen and F. Forbes. *Global implicit function theorems and the online expectation-maximisation algorithm.* 14th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03110213.

[63] B. Olivier, A. Guérin-Dugué and J.-B. Durand. *Hidden Semi-Markov Models to Segment Reading Phases from Eye Movements.* Inria Grenoble - Rhône-Alpes, 2nd Mar. 2021. URL: https://hal.inria.fr/hal-03155843.

10.3 Cited publications

[64] C. Bouveyron. 'Modélisation et classification des données de grande dimension. Application à l'analyse d'images'. PhD thesis. Université Grenoble 1, Sept. 2006. URL: http://tel.archives-ouvertes.fr/tel-00109047.

[65] P. Embrechts, C. Klüppelberg and T. Mikosh. *Modelling Extremal Events.* Vol. 33. Applications of Mathematics. Springer-Verlag, 1997.

[66] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice.* Springer Series in Statistics, Springer, 2006.

[67] S. Girard. 'Construction et apprentissage statistique de modèles auto-associatifs non-linéaires. Application à l'identification d'objets déformables en radiographie. Modélisation et classification'. PhD thesis. Université de Cery-Pontoise, Oct. 1996.

[68] K. Li. 'Sliced inverse regression for dimension reduction'. In: *Journal of the American Statistical Association* 86 (1991), pp. 316–327.

[69] B. Olivier, J.-B. Durand, A. Guérin-Dugué and M. Clausel. 'Eye-tracking data analysis using hidden semi-Markovian models to identify and characterize reading strategies'. In: *19th European Conference on Eye Movements (ECM 2017).* Wuppertal, Germany, Aug. 2017. URL: https://hal.inria.fr/hal-01671224.

[70] J. Simola, J. Salojärvi and I. Kojo. 'Using hidden Markov model to uncover processing states from eye movements in information search tasks'. In: *Cognitive Systems Research* 9.4 (Oct. 2008), pp. 237–251.