

RESEARCH CENTRE

**Grenoble - Rhône-Alpes**

IN PARTNERSHIP WITH:

**CNRS, Institut polytechnique de  
Grenoble**

2021

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

<b>Project-Team STATIFY</b>	<b>1</b>
<b>1 Team members, visitors, external collaborators</b>	<b>2</b>
<b>2 Overall objectives</b>	<b>3</b>
<b>3 Research program</b>	<b>4</b>
3.1 Mixture models	4
3.2 Graphical and Markov models	5
3.3 Functional Inference, semi- and non-parametric methods	5
3.3.1 Modelling extremal events	6
3.3.2 Level sets estimation	6
3.3.3 Dimension reduction	7
<b>4 Application domains</b>	<b>7</b>
4.1 Image Analysis	7
4.2 Biology, Environment and Medicine	7
<b>5 Social and environmental responsibility</b>	<b>8</b>
5.1 Footprint of research activities	8
5.2 Impact of research results	8
<b>6 Highlights of the year</b>	<b>8</b>
6.1 New projects	8
6.2 New responsibilities	8
6.3 Outstanding papers	8
6.4 Awards	9
<b>7 New results</b>	<b>9</b>
7.1 Mixture models	9
7.1.1 Approximate computation with surrogate posteriors	9
7.1.2 Online EM algorithm for robust clustering	9
7.1.3 Efficient Bayesian data assimilation via inverse regression	10
7.1.4 First order Sobol indices for physical models via inverse regression	10
7.1.5 The impact of asteroid shapes and topographies on their reflectance spectroscopy	10
7.1.6 Neurovascular multiparametric MRI defines epileptogenic and seizure propagation regions in experimental mesiotemporal lobe epilepsy	11
7.1.7 A non-asymptotic penalization criterion for model selection in mixture of experts models	11
7.1.8 A non-asymptotic model selection in block-diagonal mixture of polynomial experts models	12
7.1.9 Dirichlet process mixtures under affine transformations of the data	12
7.1.10 Joint supervised classification and reconstruction of irregularly sampled satellite image times series	12
7.2 Semi and non-parametric methods	13
7.2.1 Simulation of extreme events with ReLU neural networks	13
7.2.2 Estimation of extreme risk measures	13
7.2.3 Conditional extremal events	14
7.2.4 Estimation of the variability in the distribution tail	15
7.2.5 Estimation of multivariate risk measures	15
7.2.6 Dimension reduction for extremes	15
7.2.7 Dimension reduction with Sliced Inverse Regression	15
7.2.8 Diagnosing convergence of Markov chain Monte Carlo	16
7.2.9 Approximations of Bayesian nonparametric models	16
7.2.10 Applications of semi and non-parametric methods in ecology and genomics	16

7.2.11	Theoretical Analysis of Principal Components in an Umbrella Model of Intraspecific Evolution . . . . .	17
7.2.12	Adaptive potential of <i>Coffea canephora</i> from Uganda in response to climate change . . . . .	17
7.3	Graphical and Markov models . . . . .	18
7.3.1	Subtle anomaly detection in MRI brain scans: Application to biomarkers extraction in patients with <i>de novo</i> Parkinson's disease . . . . .	18
7.3.2	Leveraging 3D information in unsupervised brain MRI segmentation . . . . .	18
7.3.3	Fast Uncertainty Quantification for Deep Learning-based MR Brain Segmentation . . . . .	19
7.3.4	Patch vs. Global Image-Based Unsupervised Anomaly Detection in MR Brain Scans of Early Parkinsonian Patients . . . . .	19
7.3.5	Automatic quantification of brain lesion volume from post-trauma MR Images . . . . .	19
7.3.6	Bayesian nonparametric models for hidden Markov random fields on non-Gaussian variables and applications . . . . .	20
7.3.7	Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia . . . . .	20
7.3.8	Hidden Markov models for the analysis of eye movements . . . . .	21
7.3.9	Assessing spatial dependencies in the effect of treatment on neurite growth . . . . .	21
7.3.10	Modelling the effects of cultivars and treatments on the structure of apple trees . . . . .	22
7.3.11	Modelling growth and branching of hazelnut trees . . . . .	22
7.3.12	Estimation of leaf area densities in tropical forests . . . . .	23
7.3.13	Splitting models for multivariate count data . . . . .	23
7.3.14	Bayesian neural networks . . . . .	24
7.3.15	Brain connectivity . . . . .	24
7.4	Inverse problems . . . . .	25
7.4.1	Hierarchical bayesian models for simulation-based inference . . . . .	25
7.4.2	Bayesian inference on a large-scale brain simulator . . . . .	25
<b>8</b>	<b>Bilateral contracts and grants with industry</b> . . . . .	<b>26</b>
8.1	Bilateral contracts with industry . . . . .	26
<b>9</b>	<b>Partnerships and cooperations</b> . . . . .	<b>26</b>
9.1	International initiatives . . . . .	26
9.1.1	Inria associate team not involved in an IIL or an international program . . . . .	26
9.1.2	Participation in other International Programs . . . . .	26
9.2	International research visitors . . . . .	27
9.2.1	Visits of international scientists . . . . .	27
9.2.2	Visits to international teams . . . . .	27
9.3	National initiatives . . . . .	27
9.3.1	Networks . . . . .	28
<b>10</b>	<b>Dissemination</b> . . . . .	<b>28</b>
10.1	Promoting scientific activities . . . . .	28
10.1.1	Scientific events: organisation . . . . .	28
10.1.2	Scientific events: selection . . . . .	28
10.1.3	Journal . . . . .	29
10.1.4	Invited talks . . . . .	29
10.1.5	Scientific expertise . . . . .	30
10.1.6	Research administration . . . . .	30
10.2	Teaching - Supervision - Juries . . . . .	30
10.2.1	Teaching . . . . .	30
10.2.2	Supervision . . . . .	30
10.2.3	Juries . . . . .	31
10.2.4	Articles and contents . . . . .	31

<b>11 Scientific production</b>	<b>32</b>
11.1 Major publications	32
11.2 Publications of the year	32
11.3 Other	38
11.4 Cited publications	38

## 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, HDR]
- Sophie Achard [CNRS, Senior Researcher, HDR]
- Julyan Arbel [Inria, Researcher, HDR]
- Pedro Luiz Coelho Rodrigues [Inria, from Oct 2021, Starting Faculty Position]
- Stephane Girard [Inria, Senior Researcher, HDR]
- Pierre Wolinski [Univ Grenoble Alpes, Starting Research Position, from Oct 2021]

### Faculty Members

- Jean-Baptiste Durand [Institut polytechnique de Grenoble, Associate Professor]
- Jonathan El-Methni [Université de Paris, Associate Professor]
- Olivier Francois [Institut polytechnique de Grenoble, Professor, from Sep 2021]

### Post-Doctoral Fellow

- Pascal Dkengne Sielenou [Inria, until Feb 2021]

### PhD Students

- Louise Alamichel [Univ Grenoble Alpes, from Oct 2021]
- Yuchen Bai [Univ Grenoble Alpes, from Oct 2021]
- Meryem Bousebata [Univ Grenoble Alpes]
- Daria Bystrova [Univ Grenoble Alpes]
- Lucrezia Carboni [Univ Grenoble Alpes]
- Alexandre Constantin [Univ Grenoble Alpes]
- Benoit Kugler [Univ Grenoble Alpes]
- Hana Lbath [Univ Grenoble Alpes]
- Julia Linhart [Université Paris-Saclay, Nov 2021, Inria Saclay]
- Minh Tri Lê [Invensens]
- Theo Moins [Inria]
- Giovanni Poggiato [Univ Grenoble Alpes]
- Mariia Vladimirova [Inria]

### Technical Staff

- Pascal Dkengne Sielenou [Inria, Engineer, from Mar 2021]
- Antoine Lesieur [Inria, Engineer, from Apr 2021 until Sep 2021]

## Interns and Apprentices

- Louise Alamichel [Univ Grenoble Alpes, from Apr 2021 until Jul 2021]
- Mahtab Khademalhosseini [Inria, from Feb 2021 until Jul 2021]
- Khalil Leachouri [Inria, from Jun 2021 until Aug 2021]
- Hichem Saghrouni [Ecole normale supérieure Paris-Saclay, from Aug 2021 until Sep 2021]

## Administrative Assistant

- Geraldine Christin [Inria, until Aug 2021]

## Visiting Scientist

- Trung Tin Nguyen [Univ de Caen Basse-Normandie, Jan 2021]

## 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, Olivier Francois, Daria Bystrova, Giovanni Poggiato, Benoit Kugler, Alexandre Constantin, Louise Alamichel.

**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,  $\theta$  being the parameter, possibly multi-dimensional, usually unknown and to be estimated. We consider cases where the data naturally divides into observed data  $y = \{y_1, \dots, y_n\}$  and unobserved or missing data  $z = \{z_1, \dots, z_n\}$ . 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,

$$f(y_i; \theta) = \sum_{k=1}^K P(z_i = k; \theta) f(y_i | z_i; \theta). \quad (1)$$

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 Graphical and Markov models

**Participants:** Jean-Baptiste Durand, Florence Forbes, Julyan Arbel, Sophie Achard, Olivier Francois, Mariia Vladimirova, Lucrezia Carboni, Hana Lbath, Minh-tri Le, Yuchen Bai.

**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, Pedro Coelho Rodrigues, Pascal Dkengne Sielenou, Meryem Bousebata, Theo Moins, Pierre Wolinski, 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 3.3.2). Such non-parametric methods have become the cornerstone when dealing with functional data [92]. 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 3.3.3). 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 [94] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [91], which is based on the modelling of distribution tails (see section 3.3.1). 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.

### 3.3.1 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  $X_{1,n} \leq \dots \leq X_{n,n}$  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:

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

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

$$\frac{\ell(tx)}{\ell(x)} \rightarrow 1 \text{ as } x \rightarrow \infty \quad (3)$$

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  $\rho$  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:

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

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.

### 3.3.2 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.

### 3.3.3 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 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 [93]. 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 [90]. 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 [94].

## 4 Application domains

### 4.1 Image Analysis

**Participants:** Florence Forbes, Jean-Baptiste Durand, Stephane Girard, Pedro Coelho Rodrigues, 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 compétitivité" project I-VR, we worked on images of PC Boards.

### 4.2 Biology, Environment and Medicine

**Participants:** Florence Forbes, Stephane Girard, Jean-Baptiste Durand, Julyan Arbel, Sophie Achard, Pedro Coelho Rodrigues, Olivier Francois, Yuchen Bai, Theo Moins, Daria Bystrova, Meryem Bousebata, Lucrezia Carboni, Hana Lbath.

A third domain of applications concerns biology and medicine. We considered the use of mixture models to identify biomarkers. 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 Social and environmental responsibility

### 5.1 Footprint of research activities

The footprint of our research activities has not been assessed yet. Most of the team members have validated the "charte d'éco-responsabilité" written by a working group from Laboratoire Jean Kuntzmann, which should have practical implications in the near future.

### 5.2 Impact of research results

A lot of our developments are motivated by and target applications in medicine and environmental sciences. As such they have a social impact with a better handling and treatment of patients, in particular with brain diseases or disorders. On the environmental side, our work has an impact on geoscience-related decision making with e.g. extreme events risk analysis, planetary science studies and tools to assess biodiversity markers. However, how to truly measure and report this impact in practice is another question we have not really addressed yet.

## 6 Highlights of the year

### 6.1 New projects

- Julyan Arbel is coPi of the **Bayes-Duality** project launched with a funding of \$2.76 millions by Japan JST - French ANR for a total of 5 years starting in October 2021. The goal is to develop a new learning paradigm for Artificial Intelligence that learns like humans in an adaptive, robust, and continuous fashion. On the Japan side the project is led by Mohammad Emtiyaz Khan as the research director, and Kenichi Bannai and Rio Yokota as the two Co-PIs.
- A new ANR project entitled RADIO-AID for "Radiation induced neurotoxicity assessed by ST modeling and AI after brain radiotherapy" coordinated by S.Ancelet from IRSN has been granted for 4 years and will start in April 2022. It involves Statify, Grenoble Institute of Neurosciences, Pixyl, ICANS (Strasbourg), APHP, ICM and ENS P.Saclay.
- A new ANR project entitled HSMM-INCA for "Hidden Semi Markov Models: INference, Control and Applications" coordinated by N. Peyrard from INRAE has been granted for 4 years and will start in April 2022. It involves Statify, MIAT at INRAE Toulouse, Laboratoire de Mathématiques Appliquées de Compiègne, Laboratoire de Mathématiques Raphaël Salem (in Rouen) and Institut Montpelliérain Alexander Grothendieck.
- Statify is involved in a starting Inria project (Defi) called **ROAD-AI** with Cerema and granted for 4 years starting in October 2021. The goal is to develop advanced data analysis tools for diagnosing and monitoring road and road-related infrastructures. The Inria teams involved are FUN, Modal, Statify, Titane, Acentauri and Coati. On the Cerema side Statify will mainly work with the ENDSUM team. The PhD of Jacopo Iollo starting in January 2022 is in this context.

### 6.2 New responsibilities

Sophie Achard, since Nov. 2020, has been elected as the head of **Pole MSTIC** (with Jean-Paul Jamont, Karine Altisen and Christine Lescop) at University of Grenoble.

Julyan Arbel has been elected in November 2021 to the **board** of ISBA, the International Society for Bayesian Analysis.

Florence Forbes is deputy head of science for the Grenoble center since July 2021.

### 6.3 Outstanding papers

Stéphane Girard, Gilles Stupfler and Antoine Usseglio-Carleve got a paper published in the Annals of Statistics, entitled "Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models" [22].

## 6.4 Awards

Daria Bystrova received the best poster award at the international conference "End-to-end Bayesian learning" at CIRM in October 2021, for her work with Julyan Arbel and Mario Beraha on Bayesian block-diagonal graphical models via the Fiedler prior.

## 7 New results

### 7.1 Mixture models

#### 7.1.1 Approximate computation with surrogate posteriors

**Participants:** Florence Forbes, Julyan Arbel.

**Joint work with:** Hien Nguyen, La Trobe University Melbourne Australia and Trung Tin Nguyen 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 quality of approximations. The choice between discrepancies is an active research topic, which has mainly considered data discrepancies requiring samples of observations or distances between summary statistics. 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 L2 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 setting and can also be used as an alternative to sample-based approaches. 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. A remarkable feature of our approach is that it can be equally applied to settings where a sample of i.i.d. observations is available and to settings where a single observation is available, as a vector of measures, a time series realization or a data set reduced to a vector of summary statistics. Details can be found in [76].

#### 7.1.2 Online EM algorithm for robust clustering

**Participants:** Florence Forbes.

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

A popular way to approach clustering tasks is via a parametric mixture model. The vast majority of the work on such mixtures has been based on Gaussian mixture models. However, in some applications the tails of Gaussian distributions are shorter than appropriate or parameter estimations are affected by atypical observations (outliers). To address this issue, mixtures of so-called multiple scale Student distributions have been proposed and used for clustering. In contrast to the Gaussian case, no closed-form solution exists for such mixtures but tractability is maintained via the use of the expectation-maximisation (EM) algorithm. However, such mixtures require more parameters than the standard Student or Gaussian mixtures and the EM algorithm used to estimate the mixture parameters involves more complex numerical optimizations. Consequently, when the number of samples to be clustered becomes large, applying EM on the whole data set (Batch EM) may become costly both in terms of time

and memory requirements. A natural approach to bypass this issue is to consider an online version of the algorithm, that can incorporate the samples incrementally or in mini batches. In this work, we proposed to design a tractable online EM for mixtures of multiple scale Student distributions in order to use it then to detect subtle brain anomalies from MR brain scans for patients suffering from Parkinson disease. The application to Parkinson disease will be carried out jointly with Grenoble Institute of Neuroscience.

### 7.1.3 Efficient Bayesian data assimilation via inverse regression

**Participants:** Florence Forbes, Benoit Kugler.

**Joint work with:** Sylvain Douté from Institut de Planétologie et d'Astrophysique de Grenoble (IPAG) and Michel Gay from Gipsa-lab.

We propose a Bayesian approach to data assimilation problems, involving two steps. We first approximate the forward physical model with a parametric invertible model as described in [23], and we then use its properties to leverage the availability of a priori information. This approach is particularly suitable when a large number of inversions has to be performed. We illustrate the proposed methodology on a multilayer snowpack model. Details can be found in [53].

### 7.1.4 First order Sobol indices for physical models via inverse regression

**Participants:** Florence Forbes, Benoit Kugler.

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

In a bayesian inverse problem context, we aim at performing sensitivity analysis to help understand and adjust the physical model. To do so, we introduce indicators inspired by Sobol indices but focused on the inverse model. Since this inverse model is not generally available in closed form, we propose to use a parametric surrogate model to approximate it. The parameters of this model may be estimated via standard EM inference. Then we can exploit its tractable form and perform Monte-Carlo integration to efficiently estimate these pseudo Sobol indices. Details can be found in [52].

### 7.1.5 The impact of asteroid shapes and topographies on their reflectance spectroscopy

**Participants:** Florence Forbes, Benoit Kugler.

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

Our work on fast Bayesian inversion [23] can be used in planetary science. Here we report the comparison between unresolved reflectance spectroscopy of Solar System small bodies and laboratory measurements on reference surfaces. We measure the bidirectional reflectance spectroscopy of a powder of howardite and a sublimation residue composed of a Ceres analogue. The spectra are then inverted using the Hapke semi-empirical physical model and the MRTLS parametric model to be able to simulate the reflectance of the surfaces under any geometrical configuration needed. We note that both models enable an accurate rendering of the reflectance spectroscopy, but the MRTLS model adds less noise on the spectra compared to the Hapke model. Using the parameters resulting from the inversions, we simulate two

spherical bodies and the small bodies (1)Ceres and (4)Vesta whose surfaces are homogeneously covered with the Ceres analogue and powder of howardite respectively. We then simulate various scenarios of illumination and spectroscopic observations, i.e. spot-pointing and fly-bys, of these small bodies for phases angles between  $6^\circ$  and  $135^\circ$ . The unresolved reflectance spectroscopy of the simulated bodies is retrieved from the resulting images, and compared to the reflectance spectroscopy of the reference surface measured in the laboratory. Our results show that the photometric phase curves of the simulated bodies are different from the reference surfaces because of the variations of the local incidence and emergence angles due to the shape and topography of the surface. At low phase angle, the simulated bodies are brighter than the reference surfaces, with lower spectral slope and shallower absorption bands. We observe the maximum differences at wide phase angles with the various simulated observations of (4)Vesta due to its high surface topography. Finally, we highlight the differences in the spectral parameters derived from the unresolved observations at  $30^\circ$  with laboratory measurements acquired under a single geometrical configuration. Details can be found in [29].

#### 7.1.6 Neurovascular multiparametric MRI defines epileptogenic and seizure propagation regions in experimental mesiotemporal lobe epilepsy

**Participants:** Florence Forbes, Fabien Boux.

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

Improving the identification of the epileptogenic zone and associated seizure-spreading regions represents a significant challenge. Innovative brain-imaging modalities tracking neurovascular dynamics during seizures may provide new disease biomarkers. With use of a multi-parametric magnetic resonance imaging (MRI) analysis at 9.4 Tesla, we examined, elaborated, and combined multiple cellular and cerebrovascular MRI read-outs as imaging biomarkers of the epileptogenic and seizure-propagating regions. Analyses were performed in an experimental model of mesial temporal lobe epilepsy (MTLE) generated by unilateral intra-hippocampal injection of kainic acid (KA). Combining multi-parametric MRI acquisition and machine-learning [5] analyses delivers specific imaging identifiers to segregate the epileptogenic from the contralateral seizure-spreading hippocampi in experimental MTLE. The potential clinical value of our findings is critically discussed. More details can be found in [14]

#### 7.1.7 A non-asymptotic penalization criterion for model selection in mixture of experts models

**Participants:** Florence Forbes.

**Joint work with:** Tin Trung Nguyen and Faicel Chamroukhi, University of Caen and Hien Nguyen, La Trobe University Melbourne Australia.

Mixture of experts (MoE) is a popular class of models in statistics and machine learning that has sustained attention over the years, due to its flexibility and effectiveness. We consider the Gaussian-gated localized MoE (GLoME) regression model for modeling heterogeneous data. This model poses challenging questions with respect to the statistical estimation and model selection problems, including feature selection, both from the computational and theoretical points of view. We study the problem of estimating the number of components of the GLoME model, in a penalized maximum likelihood estimation framework. We provide a lower bound on the penalty that ensures a weak oracle inequality is satisfied by our estimator. To support our theoretical result, we perform numerical experiments on simulated and real data, which illustrate the performance of our finite-sample oracle inequality. Details can be found in [84].

### 7.1.8 A non-asymptotic model selection in block-diagonal mixture of polynomial experts models

**Participants:** Florence Forbes.

**Joint work with:** Tin Trung Nguyen and Faicel Chamroukhi, University of Caen and Hien Nguyen, La Trobe University Melbourne Australia.

Model selection via penalized likelihood type criteria is a standard task in many statistical inference and machine learning problems. It has led to deriving criteria with asymptotic consistency results and an increasing emphasis on introducing non-asymptotic criteria. We focus on the problem of modeling non-linear relationships in regression data with potential hidden graph-structured interactions between the high-dimensional predictors, within the mixture of experts modeling framework. In order to deal with such a complex situation, we investigate a block-diagonal localized mixture of polynomial experts (BLoMPE) regression model, which is constructed upon an inverse regression and block-diagonal structures of the Gaussian expert covariance matrices. We introduce a penalized maximum likelihood selection criterion to estimate the unknown conditional density of the regression model. This model selection criterion allows us to handle the challenging problem of inferring the number of mixture components, the degree of polynomial mean functions, and the hidden block-diagonal structures of the covariance matrices, which reduces the number of parameters to be estimated and leads to a trade-off between complexity and sparsity in the model. In particular, we provide a strong theoretical guarantee: a finite-sample oracle inequality satisfied by the penalized maximum likelihood estimator with a Jensen-Kullback-Leibler type loss, to support the introduced non-asymptotic model selection criterion. The penalty shape of this criterion depends on the complexity of the considered random subcollection of BLoMPE models, including the relevant graph structures, the degree of polynomial mean functions, and the number of mixture components. Details can be found in [83].

### 7.1.9 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 [12], 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.

### 7.1.10 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 [16]. An extension of this method taking into account of spectral correlation has been developed, it is submitted for publication [73].

## 7.2 Semi and non-parametric methods

### 7.2.1 Simulation of extreme events with ReLU neural networks

**Participants:** Stephane Girard.

**Joint work with:** M. Allouche and E. Gobet (CMAP, Ecole Polytechnique)

Feedforward neural networks based on Rectified linear units (ReLU) cannot efficiently approximate quantile functions which are not bounded, especially in the case of heavy-tailed distributions. We thus propose a new parametrization for the generator of a Generative adversarial network (GAN) adapted to this framework, basing on extreme-value theory. We provide an analysis of the uniform error between the extreme quantile and its GAN approximation. It appears that the rate of convergence of the error is mainly driven by the second-order parameter of the data distribution. The above results are illustrated on simulated data and real financial data, they are submitted for publication [67].

A similar investigation has been conducted to simulate fractional Brownian motion with ReLU neural networks [66].

### 7.2.2 Estimation of extreme risk measures

**Participants:** Stephane Girard, Antoine Usseglio Carleve.

**Joint work with:** A. Daouia (Univ. Toulouse), J. El-Methni (Univ. Paris), L. Gardes (Univ. Strasbourg) and G. Stupfler (Ensaï).

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  $\alpha$ -quantile of the loss distribution. Weissman extrapolation device for estimating extreme quantiles (when  $\alpha \rightarrow 0$ ) from heavy-tailed distributions is based on two estimators: an order statistic to estimate an intermediate quantile and an estimator of the tail-index. The common practice is to select the same intermediate sequence for both estimators. In [65], we show how an adapted choice of two different intermediate sequences leads to a reduction of the asymptotic bias associated with the resulting refined Weissman estimator. This new bias reduction method is fully automatic and does not involve the selection of extra parameters. Our approach is compared to other bias reduced estimators of extreme quantiles both on simulated and real data.

The Value-at-Risk however suffers from several weaknesses. First, it provides us only with a pointwise information:  $\text{VaR}(\alpha)$  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  $\alpha$ -quantile of the loss distribution. This risk measure thus takes into account the whole information contained in the upper tail of the distribution.

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 [17].

Currently available estimators of extreme expectiles are typically biased and hence may show poor finite-sample performance even in fairly large samples. In [77], 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.

### 7.2.3 Conditional extremal events

**Participants:** Stephane Girard, Antoine Usseglio Carleve.

**Joint work with:** G. Stupfler (Ensa).

As explained in Paragraph 7.2.2, 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.

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 [20].

Quantiles and expectiles belong to the wider family of  $L^p$ -quantiles. We propose in [62] to construct kernel estimators of extreme conditional  $L^p$ -quantiles. We study their asymptotic properties in the context of conditional heavy-tailed distributions and we show through a simulation study that taking  $p \in (1, 2)$  may allow to recover extreme conditional quantiles and expectiles accurately. Our estimators are also showcased on a real insurance data set.

In [22], 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.

#### 7.2.4 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 [18].

#### 7.2.5 Estimation of multivariate risk measures

**Participants:** Julyan Arbel, Stephane Girard, Antoine Usseglio Carleve.

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

Expectiles form a family of risk measures that have recently gained interest over the more common value-at-risk or return levels, primarily due to their capability to be determined by the probabilities of tail values and magnitudes of realisations at once. However, a prevalent and ongoing challenge of expectile inference is the problem of uncertainty quantification, which is especially critical in sensitive applications, such as in medical, environmental or engineering tasks. In [68], we address this issue by developing a novel distribution, termed the multivariate expectilebased distribution (MED), that possesses an expectile as a closed-form parameter. Desirable properties of the distribution, such as log-concavity, make it an excellent fitting distribution in multivariate applications. Maximum likelihood estimation and Bayesian inference algorithms are described. Simulated examples and applications to expectile and mode estimation illustrate the usefulness of the MED for uncertainty quantification.

#### 7.2.6 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 [70].

#### 7.2.7 Dimension reduction with Sliced Inverse Regression

**Participants:** Stephane Girard.

**Joint work with:** H. Lorenzo and J. Saracco (Inria Bordeaux Sud-Ouest).

Since its introduction in the early 90's, the Sliced Inverse Regression (SIR) methodology has evolved adapting to increasingly complex data sets in contexts combining linear dimension reduction with non linear regression. The assumption of dependence of the response variable with respect to only a few linear combinations of the covariates makes it appealing for many computational and real data application aspects. In [19], we propose an overview of the most active research directions in SIR modeling from multivariate regression models to regularization and variable selection.

### 7.2.8 Diagnosing convergence of Markov chain Monte Carlo

**Participants:** Julyan Arbel, Theo Moins, Stephane Girard.

**Joint work with:** A. Dutfoy (EDF R&D).

Diagnosing convergence of Markov chain Monte Carlo (MCMC) is crucial in Bayesian analysis. Among the most popular methods, the potential scale reduction factor (commonly named  $\hat{R}$ ) is an indicator that monitors the convergence of output chains to a stationary distribution, based on a comparison of the between- and within-variance of the chains. Several improvements have been suggested since its introduction in the 90'ss. We analyse some properties of the theoretical value  $R$  associated to  $\hat{R}$  in the case of a localized version that focuses on quantiles of the distribution. This leads to proposing a new indicator, which is shown to allow both for localizing the MCMC convergence in different quantiles of the distribution, and at the same time for handling some convergence issues not detected by other  $\hat{R}$  versions. A preliminary version of this work is published in [25].

### 7.2.9 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)

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 [31], 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.

### 7.2.10 Applications of semi and non-parametric methods in ecology and genomics

**Participant:** Julyan Arbel, Daria Bystrova, Giovanni Poggiato.

**Joint work with:** Billur Bektaş, Tamara Münkemüller, and Wilfried Thuiller, LECA - Laboratoire d'Ecologie Alpine, James S Clark from Nicholas School of the Environment, Duke University, USA, Alessandra Guglielmi, POLIMI - Dipartimento di Matematica - POLIMI, Politecnico di Milano.

Explaining and modelling species communities is more than ever a central goal of ecology. Recently, joint species distribution models (JSDMs), which extend species distribution models (SDMs) by considering correlations among species, have been proposed to improve species community analyses and rare species predictions while potentially inferring species interactions. Here, we illustrate the mathematical links between SDMs and JSDMs and their ecological implications and demonstrate that JSDMs, just like SDMs, cannot separate environmental effects from biotic interactions. We provide a guide to the conditions under which JSDMs are (or are not) preferable to SDMs for species community modelling. More generally, we call for a better uptake and clarification of novel statistical developments in the field of biodiversity modelling. This work is published in [28].

In [15], 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. A book chapter describing latent factor models as a tool for dimension reduction in joint species distribution models is also available [72].

### 7.2.11 Theoretical Analysis of Principal Components in an Umbrella Model of Intraspecific Evolution

**Participant:** Olivier Francois.

**Joint work with:** Maxime Estavoyer

We proposed a mathematical theory for the observation of sinusoidal patterns in the principal component analysis of human genetic data when evolution is based on a hierarchy of splits from an ancestral population. Our results provided detailed mathematical descriptions of eigenvalues and principal components of sampled genomic sequences under the split model. Removing variants uniquely represented in the sample, the PCs are defined as cosine functions of increasing periodicity, reproducing wave-like patterns observed in equilibrium isolation-by-distance models. Including rare variants in the analysis, the PCs corresponding to the largest eigenvalues exhibit complex wave shapes. Genomic data related to the peopling of the Americas are reanalyzed in the light of our new model.

### 7.2.12 Adaptive potential of *Coffea canephora* from Uganda in response to climate change

**Participant:** Olivier Francois.

In collaboration with IRD, we developed statistical methods to evaluate genetic maladaptation in plant populations facing rapid change climate. We applied the new measures in a study of 207 coffee (*Coffea canephora*) trees from seven forests along climatic gradients in Uganda. The estimates of genetic vulnerability allowed us to target new elite parents from the northernmost zone of Uganda for future breeding programs.

### 7.3 Graphical and Markov models

#### 7.3.1 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.

#### 7.3.2 Leveraging 3D information in unsupervised brain MRI segmentation

**Participants:** Florence Forbes.

**Joint work with:** Benjamin Lambert, Maxime Louis, Senan Doyle, Alan Tucholka from Pixyl and Michel Dojat from Grenoble Institute of Neurosciences.

Automatic segmentation of brain abnormalities is challenging, as they vary considerably from one pathology to another. Current methods are supervised and require numerous annotated images for each pathology, a strenuous task. To tackle anatomical variability, Unsupervised Anomaly Detection (UAD) methods are proposed, detecting anomalies as outliers of a healthy model learned using a Variational Autoencoder (VAE). Previous work on UAD adopted 2D approaches, meaning that MRIs are processed as a collection of independent slices. Yet, it does not fully exploit the spatial information contained in MRI. Here, we propose to perform UAD in a 3D fashion and compare 2D and 3D VAEs. As a side contribution, we present a new loss function guarantying a robust training. Learning is performed

using a multicentric dataset of healthy brain MRIs, and segmentation performances are estimated on White-Matter Hyperintensities and tumors lesions. Experiments demonstrate the interest of 3D methods which outperform their 2D counterparts. Details can be found in [33].

### 7.3.3 Fast Uncertainty Quantification for Deep Learning-based MR Brain Segmentation

**Participants:** Florence Forbes.

**Joint work with:** Benjamin Lambert, Senan Doyle, Alan Tucholka from Pixyl and Michel Dojat from Grenoble Institute of Neurosciences.

Quantifying the uncertainty attached to Deep Learning models predictions can help their interpretation, and thus their acceptance in critical fields. Yet, current standard approaches rely on multi-steps approaches, increasing the inference time and memory cost. In clinical routine, the automated prediction has to integrate into the clinical consultation timeframe, raising the need for faster and more efficient uncertainty quantification methods. In this work, we propose a novel model, named as BEHT, and evaluate it on an automated segmentation task of White-Matter Hyperintensities from T2-weighted FLAIR MRI sequences of Multiple-Sclerosis (MS) patients. We demonstrate that this approach outputs predictive uncertainty much faster than the state-of-the-art Monte Carlo Dropout approach, with a similar-and even slightly better-accuracy. Interestingly, our approach distinguishes 2 distinct sources of uncertainties, namely aleatoric and epistemic uncertainties. Details can be found in [54].

### 7.3.4 Patch vs. Global Image-Based Unsupervised Anomaly Detection in MR Brain Scans of Early Parkinsonian Patients

**Participants:** Florence Forbes, Veronica Munoz-Ramirez.

**Joint work with:** Michel Dojat from Grenoble Institute of Neurosciences, Carole Lartizien, Nicolas Pinon from Creatis.

Although neural networks have proven very successful in a number of medical image analysis applications, their use remains difficult when targeting subtle tasks such as the identification of barely visible brain lesions, especially given the lack of annotated datasets. Good candidate approaches are patch-based unsupervised pipelines which have both the advantage to increase the number of input data and to capture local and fine anomaly patterns distributed in the image, while potential inconveniences are the loss of global structural information. We illustrate this trade-off on Parkinson's disease (PD) anomaly detection comparing the performance of two anomaly detection models based on a spatial auto-encoder (AE) and an adaptation of a patch-fed siamese auto-encoder (SAE). On average, the SAE model performs better, showing that patches may indeed be advantageous. More details can be found in [35].

### 7.3.5 Automatic quantification of brain lesion volume from post-trauma MR Images

**Participants:** Florence Forbes.

**Joint work with:** Senan Doyle, Alan Tucholka from Pixyl and Michel Dojat from Grenoble Institute of Neurosciences.

The determination of the volume of brain lesions after trauma is challenging. Manual delineation is observer-dependent and time-consuming which inhibits the practice in clinical routine. We propose and

evaluate an automated atlas-based quantification procedure (AQP) based on the detection of abnormal mean diffusivity (MD) values computed from diffusion-weighted MR images. We measured the performance of AQP versus manual delineation consensus by independent raters in two series of experiments: i) realistic trauma phantoms ( $n=5$ ) where abnormal MD values were assigned to healthy brain images according to the intensity, form and location of lesion observed in real TBI cases; ii) severe TBI patients ( $n=12$  patients) who underwent MR imaging within 10 days after injury. In realistic trauma phantoms, no statistical difference in Dice similarity coefficient, precision and brain lesion volumes was found between AQP, the rater consensus and the ground truth lesion delineations. Similar findings were obtained when comparing AQP and manual annotations for TBI patients. The intra-class correlation coefficient between AQP and manual delineation was 0.70 in realistic phantoms and 0.92 in TBI patients. The volume of brain lesions detected in TBI patients was 59 ml (19-84 ml) (median; 25-75th centiles). Our results indicate that an automatic quantification procedure could accurately determine with accuracy the volume of brain lesions after trauma. This presents an opportunity to support the individualized management of severe TBI patients. Details can be found in [24].

### 7.3.6 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 [11] 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 [75]. 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.

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.

### 7.3.7 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 [75].

### 7.3.8 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 through eye-movements. Our approach was inspired by Simola *et al.* (2008) [95], but we used hidden semi-Markov models for better characterization of phase length distributions (Olivier *et al.*, 2021) [85]. 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.

The results of this study have been partially included in a dissemination work at Inria Interstice journal.

### 7.3.9 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 7.3.7. 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.

Progresses have been made in the image processing step: segmentation of neurons and their neurites together with computation of neurite lengths.

### 7.3.10 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, Sandra Plancace and Nathalie Peyrard, INRAE, MIAT, Toulouse, 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 7.3.8.

It is now investigated how the HSMC parameters depend on cultivars and treatments using generalized linear mixed models (GLMMs). Fixed effects represent how conditional probabilities of observations given states, state transitions and dwelling times depend on cultivars and treatments, while random effects represent variability and dependencies due to repeated measurements on a same tree. Bayesian estimation of parameters with MCMC highlighted some difficulty to reach convergence, inducing high variability of a Bayesian information criterion (DIC) for model comparison. The ability of DIC and GLMMs to assess the effects of the different factors is now investigated using permutation tests, to compare the DIC values obtained on real data to those obtained by removing the effects using permutations of observations.

### 7.3.11 Modelling growth and branching of hazelnut trees

**Participants:** Jean-Baptiste Durand.

**Joint work with:** Evelyne Costes, INRAE, AGAP, Montpellier, France, Frédéric Boudon, CIRAD, AGAP, Montpellier, France, Francesca Grisafi and Sergio Tombesi, Università Cattolica del Sacro Cuore, Piacenza,

Italy.

This work aims at developing a functional-structural plant model to represent the main components in tree growth for hazelnut trees. Data were collected over two years at annual shoot and metamer scales. Exploratory analyses were achieved to represent the structure of the lateral production of buds, depending on their position on the shoot and accounting for possible dependencies of the numbers of latent vs. developing buds of different types: blind, vegetative, mixed (i.e., bearing female reproductive organs with potential nuts) and catkins (i.e., bearing male reproductive organs).

Future work will involve building a probabilistic model for these dependencies, which will be a building block for simulation models at whole tree, shoot and metamer scales. The outputs of the model will have to be validated using independent validation data and different metrics combining structural and functional components of the data (to be defined).

### 7.3.12 Estimation of leaf area densities in tropical forests

**Participants:** Jean-Baptiste Durand, Florence Forbes, Yuchen Bai.

**Joint work with:** Grégoire Vincent, IRD, AMAP, Montpellier, France.

Covering just 7% of the Earth's land surface, tropical forests play a disproportionate role in the biosphere: they store about 25% of the terrestrial carbon and contribute to over a third of the global terrestrial productivity. They also recycle about a third of the precipitations through evapotranspiration and thus contribute to generate and maintain a humid climate regionally, with positive effects also extending well beyond the tropics. However, the seasonal variability in fluxes between tropical rainforests and atmosphere is still poorly understood. Better understanding the processes underlying flux seasonality in tropical forests is thus critical to improve our predictive ability on global biogeochemical cycles. Leaf area, one key variable controlling water efflux and carbon influx, is poorly characterized. To monitor evolutions of biomass, leaf area density (LAD) or gas exchange, aerial and terrestrial laser scanner (LiDAR) measurements have been frequently used.

The principle is, for different LiDAR shoots assumed as independent, to measure the portions of beam lengths between successive hits. Possible censoring comes from beams not being intercepted within a given voxel. Current approaches aim at connecting LAD to the distribution of beam lengths through some statistical model. Such a simplified model does not currently take into account several effects that may impact either LAD or beam lengths: heterogeneity and dependencies in the vegetation properties in different voxels on the one hand and the nature of hit material on the other hand: wood vs. leaves, leading to biases or deteriorated uncertainty in estimation.

This collaboration, supported by Y. Bai's PhD work, aims at developing machine learning methods to address these issues. The potential added-value of Hidden Markov random fields (see Section 7.3.7) to go beyond the assumption of independent voxels and taking into account spatial dependencies between them has been investigated. Current work is now focusing on applying 3D point convolution neural network to discriminate wood from leaves.

### 7.3.13 Splitting models for multivariate count data

**Participants:** Jean-Baptiste Durand.

**Joint work with:** Jean Peyhardi, Institut Montpellierain 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 [27]. 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.

#### 7.3.14 Bayesian neural networks

**Participants:** Julyan Arbel, Mariia Vladimirova, Stephane Girard.

The connection between Bayesian neural networks and Gaussian processes gained a lot of attention in the last few years, with the flagship result that hidden units converge to a Gaussian process limit when the layers width tends to infinity. Underpinning this result is the fact that hidden units become independent in the infinite-width limit. Our aim is to shed some light on hidden units dependence properties in practical finite-width Bayesian neural networks. In addition to theoretical results, we assess empirically the depth and width impacts on hidden units dependence properties. This work is published in [36].

Hidden units are proven to follow a Gaussian process limit when the layer width tends to infinity. Recent work has suggested that finite Bayesian neural networks may outperform their infinite counterparts because they adapt their internal representations flexibly. To establish solid ground for future research on finite-width neural networks, our goal is to study the prior induced on hidden units. Our main result is an accurate description of hidden units tails which shows that unit priors become heavier-tailed going deeper, thanks to the introduced notion of generalized Weibull-tail. This finding sheds light on the behavior of hidden units of finite Bayesian neural networks. This work is published in [60].

#### 7.3.15 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 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.

## 7.4 Inverse problems

### 7.4.1 Hierarchical bayesian models for simulation-based inference

**Participants:** Pedro Rodrigues, Julia Linhart.

**Joint work with:** Thomas Moreau and Alexandre Gramfort from Inria Saclay and Gilles Louppe from Université de Liège

Inferring the parameters of a stochastic model based on experimental observations is central to the scientific method. A particularly challenging setting is when the model is strongly indeterminate, i.e. when distinct sets of parameters yield identical observations. This arises in many practical situations, such as when inferring the distance and power of a radio source (is the source close and weak or far and strong?) or when estimating the amplifier gain and underlying brain activity of an electrophysiological experiment. In a recent work, we have proposed the hierarchical neural posterior estimation (HNPE), a novel method for cracking such indeterminacy by exploiting additional information conveyed by an auxiliary set of observations sharing global parameters. This method extends recent developments in simulation-based inference (SBI) based on normalizing flows to Bayesian hierarchical models. We validated HNPE quantitatively on a motivating example amenable to analytical solutions and then applied it to invert a well known non-linear model from computational neuroscience, using both simulated and real EEG data.

### 7.4.2 Bayesian inference on a large-scale brain simulator

**Participants:** Pedro Rodrigues.

**Joint work with:** Nicholas Tolley and Stephanie Jones from Brown University, Alexandre Gramfort from Inria Saclay

The Human Neocortical Neurosolver (HNN) is a framework whose foundation is a cortical column model with cell and circuit level detail designed to connect macroscale signals to meso/microcircuit level phenomena. We apply this model to study the cellular and circuit mechanisms of beta generation using local field potential (LFP) recordings from the non-human primate (NHP) motor cortex. To characterize beta producing mechanisms, we employ simulation based inference (SBI) in the HNN modeling tool. This framework leverages machine learning techniques and neural density estimators to characterize the relationship between a large space of model parameters and simulation output. In this setting, Bayesian inference can be applied to models with intractable likelihood functions (Gonçalves 2020, Papamakarios 2021). The main goal of this project is to provide a set of guidelines for scientists that wish to apply simulation-based inference to their neuroscience studies with a large-scale simulator such as HNN. This involves developing new methods for extracting summary features, checking the quality of the posterior approximation, etc. This work is mostly carried out by the Ph.D. student Nicholas Tolley from Brown University.

## 8 Bilateral contracts and grants with industry

### 8.1 Bilateral contracts with industry

**Participants:** Stephane Girard, Julyan Arbel.

**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.

## 9 Partnerships and cooperations

### 9.1 International initiatives

#### 9.1.1 Inria associate team not involved in an IIL or an international program

**Participants:** Florence Forbes, Julyan Arbel, Jean-Baptiste Durand, Stephane Girard, Benoit Kugler, TinTrung Nguyen.

LANDER stands for "Latent Analysis, Adversarial Networks, and DimEnsionality Reduction" and is an associate team with researchers in Australia that has started in 2019. Details can be found at [Lander website](#).

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.

#### 9.1.2 Participation in other International Programs

**Participants:** Florence Forbes, Julyan Arbel, Sophie Achard, Hana Lbath.

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.

Julyan Arbel is coPI of the **Bayes-Duality** project launched with a funding of \$2.76 millions by Japan JST - French ANR for a total of 5 years starting in October 2021. The goal is to develop a new learning paradigm for Artificial Intelligence that learns like humans in an adaptive, robust, and continuous fashion. The financial support for the French side is 464k euros.

## 9.2 International research visitors

### 9.2.1 Visits of international scientists

Bernardo Nipoti, Bicocca University, Milan, Italy, visited the team in June for a collaboration with Julyan Arbel on Beta two-parameter processes in Bayesian nonparametrics.

### 9.2.2 Visits to international teams

**Research stays abroad** Julyan Arbel visited the Casa Matemática Oaxaca (CMO), Mexico (November 28 - December 3) for a working group dedicated to bringing together researchers working on objective Bayesian methodology, Bayesian non-parametric methods, and machine learning. The Casa Matematica Oaxaca (CMO) in Mexico, and the Banff International Research Station for Mathematical Innovation and Discovery (BIRS) in Banff, Canada, are collaborative Canada-US-Mexico ventures that provide an environment for creative interaction as well as the exchange of ideas, knowledge, and methods within the mathematical sciences, with related disciplines and with industry. Other BIRS partners include the Institute for Advanced Study in Mathematics (Hangzhou, China) and the Institute of Mathematics at the University of Granada (Spain).

## 9.3 National initiatives

**Participants:** Jean-Baptiste Durand, Florence Forbes, Julyan Arbel, Sophie Achard, Stephane Girard, Alexandre Constantin, Meryem Bousebata, Giovanni Poggiato.

**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 cross-disciplinary project (CDP) **Risk@UGA**.

- 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.

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".

### 9.3.1 Networks

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

## 10 Dissemination

**Participants:** Sophie Achard, Pedro Coelho Rodrigues, Florence Forbes, Julyan Arbel, Jean-Baptiste Durand, Stephane Girard, Olivier Francois, Hana Lbath.

### 10.1 Promoting scientific activities

#### 10.1.1 Scientific events: organisation

##### General chair, scientific chair

- Brain connectivity networks: quality and reproducibility workshop in the context of the Complex Systems conference in October 2021 in Lyon. Details on the [workshop](#) website.

##### Member of the organizing committees

- Julyan Arbel was a member of the scientific and organizing committees of the French Mirror of the ISBA conference at CIRM in July 2021, of the Approximate Bayesian Computation meeting in April 2021, and of the Statistical Methods for Post Genomic Data analysis (SMPGD) meeting, in January 2021.

#### 10.1.2 Scientific events: selection

##### Reviewer

- Julyan Arbel has been a reviewer for the International Conference on Machine Learning (ICML), the Symposium on Advances in Approximate Bayesian Inference (AABI).



- Florence Forbes has been a reviewer for EUSIPCO 2021.

### 10.1.3 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.
- 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 reviewer for Biometrika, Journal of Royal Statistical Society - series C, Journal of Multivariate Analysis, Scandinavian Journal of Statistics, IEEE Transactions on Signal Processing, Econometrics and Statistics, and for a book to be published in CRC Press.
- Stéphane Girard has been a reviewer for JASA (Journal of the American Statistical Association) and EJS (Electronic Journal of Statistics).
- Florence Forbes has been a reviewer for Statistics & Computing, Bayesian Analysis, Australian and New Zealand journal of Statistics, Journal of Statistical Distributions and Applications.
- Pedro Rodrigues has been a reviewer for NeuroImage, IEEE Transactions on Biomedical Engineering, and Pattern Recognition.
- Jean-Baptiste Durand has been a reviewer for Ecology and Evolution and for Advances in Data Analysis and Classification.

### 10.1.4 Invited talks

- Florence Forbes has been invited as a plenary speaker to the international conference "End-to-end Bayesian Learning" at CIRM in October 2021, and to the workshop in honor of Christian Robert 60th birthday in September 2021.  
F. Forbes was also invited to give a talk in special sessions at the ABC in Svalbard international conference and at the SIAM international conference on Uncertainty Quantification respectively in April and March 2021.  
F. Forbes was then invited to give talks at the Laplace Daemon Criteo Online Seminar in March 2021, at the local Statistics department seminar at University of Grenoble and at the Grenoble Institute of Neuroscience neuroimaging team days in September 2021.
- Julyan Arbel has been invited to give a talk at the OxCSML Seminar at University of Oxford in April 2021, at the ApproxBayes team Seminar, RIKEN AIP, Tokyo, Japan in May, at Journées MAS (Modélisation Aléatoire et Statistique), France, in August, and at the Foundations of Objective Bayesian Methodology Workshop, Casa Matemática Oaxaca (CMO), Mexico, in December.
- Stéphane Girard was an invited speaker at the 14th International Conference of the ERCIM WG on Computing and Statistics [50] and the 13th International Workshop on Rare-Event Simulation [51].
- Pedro Rodrigues was invited to give a talk to the GAIA team from GIPSA-lab in December 2021.

### 10.1.5 Scientific expertise

- Florence Forbes has reviewed projects for the Swiss Personalized Health and Related Technologies (PHRT) Pioneer Imaging Projects.
- Florence Forbes is a member of the Helmholtz AI Cooperation Unit advisory committee, 2019-present.
- Florence Forbes and Sophie Achard are members of the EURASIP Technical Area Committee BISA (Biomedical Image & Signal Analytics) since January 2021 for a 3 years duration.
- Florence Forbes was a member of committees in charge of hiring professors and teaching assistants at Ecole Polytechnique, Paris, and at the Agricultural Science School in Rennes (AgroCampus ouest), and of the committee in charge of hiring Inria junior researchers for the Grenoble center.
- Stéphane Girard was a reviewer for the Hi!Paris Fellowships program 2021 Call.

### 10.1.6 Research administration

- Sophie Achard, since Nov. 2020, has been elected as the head of **Pole MSTIC** (with Jean-Paul Jamont, Karine Altisen and Christine Lescop) at University of Grenoble.
- Florence Forbes is since July 2021 Deputy head of science (DSA) for the Inria Grenoble center.
- Julyan Arbel is a member of the scientific committee of the Data Science axis of Persyval Labex.

## 10.2 Teaching - Supervision - Juries

### 10.2.1 Teaching

- Master : Stéphane Girard, *Statistique Inférentielle Avancée*, 18 ETD, M1 level, Ensimag. Grenoble-INP, France.
- Master : Stéphane Girard, *Introduction to Extreme-Value Analysis*, 27 ETD, M2 level, Univ-Grenoble Alpes (UGA), 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.

### 10.2.2 Supervision

PhD Defended: Benoit Kugler, "Massive hyperspectral images analysis by inverse regression of physical models", Florence Forbes and Sylvain Douté, Université Grenoble Alpes, Defended in July 2021.

PhD Defended: Alexandre Constantin "Analyse de séries temporelles massives d'images satellitaires : Applications à la cartographie des écosystèmes", Stéphane Girard and Mathieu Fauvel, Université Grenoble Alpes, Defended in December 2021.

PhD in progress: Mariia Vladimirova, "Prior specification for Bayesian deep learning models and regularization implications", started on October 2018, Julyan Arbel, Université Grenoble Alpes.

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: 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: 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: 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.

PhD in progress: Hana Lbath, "Advanced Spatiotemporal Statistical Models for Quantification and Estimation of Functional Connectivity", started in October 2020, supervised by Sophie Achard and Alex Petersen (Brigham Young University, Utah, USA).

PhD in progress: Lucrezia Carboni, "Graph embedding for brain connectivity", started in October 2020, supervised by Sophie Achard and Michel Dojat (GIN).

PhD in progress: Louise Alamichel. "Bayesian Nonparametric methods for complex genomic data" Inria, started in October 2021, advised by Julyan Arbel and Guillaume Kon Kam King (INRAE).

PhD in progress: Yuchen Bai, "Hierarchical Bayesian Modelling of leaf area density from UAV-lidar", started in October 2021, supervised by Jean-Baptiste Durand, Florence Forbes and Gregoire Vincent (IRD, Montpellier).

PhD in progress: Julia Linhart, "Simulation based inference with neural networks: applications to computational neuroscience", started in November 2021, supervised by Pedro Rodrigues and Alexandre Gramfort (DR Inria Saclay).

### 10.2.3 Juries

- Florence Forbes has been a reviewer for the HDR of Pierre Maurel (Rennes).
- Florence Forbes has been a member of the PhD committees of Hamza Cherkaoui (Saclay), Amelie Barbe (Lyon); and chair for the committees of Olga Permiakova (Grenoble) and Remi Souriau (Saclay).
- Florence Forbes has been a member of the intermediate PhD committee of Nicolas Pinon (Lyon) and reviewer for the Master committee of Michael Carr (QUT Brisbane Australia).
- Stéphane Girard has been a member of the PhD committee of Benoit Colange (Université Savoie-Mont Blanc).
- Stéphane Girard has been a member of the intermediate PhD committee of Valentine Bellet and Erwan Giry-Fouquet (Université de Toulouse).

### 10.2.4 Articles and contents

Sophie Achard and Jean-Baptiste Durand published an illustration of the process of statistical modeling, in the Inria journal for popularization of science (in French). The topic was illustrated through the analysis of eye movements to infer cognitive processes (see Subsection 7.3.8).

## 11 Scientific production

### 11.1 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 Analysis 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] 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* 40.7 (July 2021), pp. 1827–1837. DOI: [10.1109/TMI.2021.3066781](https://doi.org/10.1109/TMI.2021.3066781). URL: <https://hal.archives-ouvertes.fr/hal-02314026>.
- [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](https://doi.org/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](https://doi.org/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.
- [11] 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](https://doi.org/10.1007/s11222-020-09935-9). URL: <https://hal.archives-ouvertes.fr/hal-02163046>.

### 11.2 Publications of the year

#### International journals

- [12] 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](https://doi.org/10.1007/s00180-020-01013-y). URL: <https://hal.archives-ouvertes.fr/hal-01950652>.
- [13] 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* 40.7 (July 2021), pp. 1827–1837. DOI: [10.1109/TMI.2021.3066781](https://doi.org/10.1109/TMI.2021.3066781). URL: <https://hal.archives-ouvertes.fr/hal-02314026>.
- [14] F. Boux, F. Forbes, N. Collomb, E. Zub, 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* 62.5 (May 2021), pp. 1244–1255. DOI: [10.1111/epi.16886](https://doi.org/10.1111/epi.16886). URL: <https://hal.archives-ouvertes.fr/hal-03193021>.

- [15] 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* 9 (9th Mar. 2021), 601384:1–11. DOI: [10.3389/fevo.2021.601384](https://doi.org/10.3389/fevo.2021.601384). URL: <https://hal.inria.fr/hal-03151472>.
- [16] A. Constantin, M. Fauvel and S. Girard. ‘Joint Supervised Classification and Reconstruction of Irregularly Sampled Satellite Image Times Series’. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (11th May 2021), p. 4403913. DOI: [10.1109/TGRS.2021.3076667](https://doi.org/10.1109/TGRS.2021.3076667). URL: <https://hal.inria.fr/hal-02997573>.
- [17] A. Daouia, S. Girard and G. Stupfler. ‘ExpectHill estimation, extreme risk and heavy tails’. In: *Journal of Econometrics* 221.1 (Mar. 2021), pp. 97–117. DOI: [10.1016/j.jeconom.2020.02.003](https://doi.org/10.1016/j.jeconom.2020.02.003). URL: <https://hal.inria.fr/hal-01856212>.
- [18] L. Gardes and S. Girard. ‘On the estimation of the variability in the distribution tail’. In: *Test* 30 (Dec. 2021), pp. 884–907. DOI: [10.1007/s11749-021-00754-2](https://doi.org/10.1007/s11749-021-00754-2). URL: <https://hal.inria.fr/hal-02400320>.
- [19] S. Girard, H. Lorenzo and J. Saracco. ‘Advanced topics in Sliced Inverse Regression’. In: *Journal of Multivariate Analysis* 188 (2022), p. 104852. DOI: [10.1016/j.jmva.2021.104852](https://doi.org/10.1016/j.jmva.2021.104852). URL: <https://hal.inria.fr/hal-03367798>.
- [20] S. Girard, G. Stupfler and A. Usseglio-Carleve. ‘Functional estimation of extreme conditional expectiles’. In: *Econometrics and Statistics* 21 (Jan. 2022), pp. 131–158. DOI: [10.1016/j.ecosta.2021.05.006](https://doi.org/10.1016/j.ecosta.2021.05.006). URL: <https://hal.inria.fr/hal-03117547>.
- [21] S. Girard, G. Stupfler and A. Usseglio-Carleve. ‘Nonparametric extreme conditional expectile estimation’. In: *Scandinavian Journal of Statistics* 49.1 (Mar. 2022), pp. 78–115. DOI: [10.1111/sjos.12502](https://doi.org/10.1111/sjos.12502). URL: <https://hal.archives-ouvertes.fr/hal-02114255>.
- [22] S. Girard, G. C. Stupfler and A. Usseglio-Carleve. ‘Extreme Conditional Expectile Estimation in Heavy-Tailed Heteroscedastic Regression Models’. In: *Annals of Statistics* 49.6 (Dec. 2021), pp. 3358–3382. DOI: [10.1214/21-AOS2087](https://doi.org/10.1214/21-AOS2087). URL: <https://hal.archives-ouvertes.fr/hal-03306230>.
- [23] B. Kugler, F. Forbes and S. Douté. ‘Fast Bayesian Inversion for high dimensional inverse problems’. In: *Statistics and Computing* (2021). URL: <https://hal.archives-ouvertes.fr/hal-02908364>.
- [24] T. Mistral, P. Roca, C. Maggia, A. Tucholka, F. Forbes, S. Doyle, A. Krainik, D. Galanaud, E. Schmitt, S. Kremer, A. Kastler, I. Troprès, E. Barbier, J.-F. Payen and M. Dojat. ‘Automatic quantification of brain lesion volume from post-trauma MR Images’. In: *Frontiers in Neurology*. section Neurotrauma (2021). DOI: [10.1101/2021.04.24.21255599](https://doi.org/10.1101/2021.04.24.21255599). URL: <https://hal.archives-ouvertes.fr/hal-03498127>.
- [25] T. Moins, J. Arbel, A. Dutfoy and S. Girard. ‘Discussion of the paper "Rank-Normalization, Folding, and Localization: An Improved  $\hat{R}$  for Assessing Convergence of MCMC"’. In: *Bayesian Analysis* 16.2 (2021), pp. 711–712. DOI: [10.1214/20-ba1221](https://doi.org/10.1214/20-ba1221). URL: <https://hal.inria.fr/hal-03222934>.
- [26] H. D. Nguyen and F. Forbes. ‘Global implicit function theorems and the online Expectation-Maximisation algorithm’. In: *Australian and New Zealand Journal of Statistics* (2022), pp. 1–27. DOI: [10.1111/anzs.12356](https://doi.org/10.1111/anzs.12356). URL: <https://hal.archives-ouvertes.fr/hal-03110213>.
- [27] 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](https://doi.org/10.1016/j.jmva.2020.104677). URL: <https://hal.inria.fr/hal-02962877>.
- [28] 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* 36.5 (1st May 2021), pp. 391–401. DOI: [10.1016/j.tree.2021.01.002](https://doi.org/10.1016/j.tree.2021.01.002). URL: <https://hal.archives-ouvertes.fr/hal-03153558>.
- [29] S. M. Potin, S. Douté, B. Kugler and F. Forbes. ‘The impact of asteroid shapes and topographies on their reflectance spectroscopy’. In: *Icarus* 376 (Apr. 2022), 114806:1–24. DOI: [10.1016/j.icarus.2021.114806](https://doi.org/10.1016/j.icarus.2021.114806). URL: <https://hal.archives-ouvertes.fr/hal-03500909>.

- [30] 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](https://doi.org/10.1162/netn_a_00176). URL: <https://hal.archives-ouvertes.fr/hal-03165916>.

### International peer-reviewed conferences

- [31] 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>.
- [32] L. Carboni, S. Achard and M. Dojat. ‘Network embedding for brain connectivity’. In: ISBI 2021 - International Symposium on Biomedical Imaging. Nice / Virtual, France, 13th Apr. 2021, pp. 1–4. URL: <https://hal.archives-ouvertes.fr/hal-03273946>.
- [33] B. Lambert, M. Louis, S. Doyle, F. Forbes, M. Dojat and A. Tucholka. ‘Leveraging 3D information in unsupervised brain MRI segmentation’. In: ISBI 2021 - 18th International Symposium on Biomedical Imaging. Nice / Virtual, France, 13th Apr. 2021, pp. 1–4. URL: <https://hal.archives-ouvertes.fr/hal-03273953>.
- [34] H. Lbath, A. Bonifati and R. Harmer. ‘Schema Inference for Property Graphs’. In: EDBT 2021 - 24th International Conference on Extending Database Technology. EDBT. Nicosia, Cyprus, Mar. 2021, pp. 499–504. DOI: [10.5441/002/edbt.2021.58](https://doi.org/10.5441/002/edbt.2021.58). URL: <https://hal.archives-ouvertes.fr/hal-03361480>.
- [35] V. Muñoz-Ramírez, N. Pinon, F. Forbes, C. Lartizien and M. Dojat. ‘Patch vs. Global Image-Based Unsupervised Anomaly Detection in MR Brain Scans of Early Parkinsonian Patients’. In: MLCN 2021 - 4th International Workshop in Machine Learning in Clinical Neuroimaging. Vol. 13001. Lecture Notes in Computer Science. Strasbourg, France: Springer International Publishing, 21st Sept. 2021, pp. 34–43. DOI: [10.1007/978-3-030-87586-2\\_4](https://doi.org/10.1007/978-3-030-87586-2_4). URL: <https://hal.archives-ouvertes.fr/hal-03397081>.
- [36] M. Vladimirova, J. Arbel and S. Girard. ‘Dependence between Bayesian neural network units’. In: BDL 2021 - Workshop. Bayesian Deep Learning NeurIPS. Montreal, Canada, 14th Dec. 2021, pp. 1–9. URL: <https://hal.inria.fr/hal-03449211>.

### National peer-reviewed Conferences

- [37] M. Allouche, S. Girard and E. Gobet. ‘On the approximation of extreme quantiles with neural networks’. In: SFdS 2021 - 52èmes Journées de Statistique de la Société Française de Statistique. Nice, France, 7th June 2021, pp. 1–5. URL: <https://hal.inria.fr/hal-03268702>.
- [38] M. Bousebata, G. Enjolras and S. Girard. ‘Single-index Extreme-PLS regression’. In: JDS 2021 - 52èmes Journées de Statistique organisées par la Société Française de Statistique (SFdS). Nice / Virtual, France, 7th June 2021, pp. 1–6. URL: <https://hal.inria.fr/hal-03267334>.
- [39] T. Moins, J. Arbel, A. Dutfoy and S. Girard. ‘On Reparameterisations of the Poisson Process Model for Extremes in a Bayesian Framework’. In: JDS 2021 - 52èmes Journées de Statistique de la Société Française de Statistique (SFdS). Nice / Virtual, France, 7th June 2021, pp. 1–6. URL: <https://hal.inria.fr/hal-03264261>.

### Conferences without proceedings

- [40] M. Allouche, S. Girard and E. Gobet. ‘Generative model for fbm with deep ReLU neural networks’. In: Bernoulli-IMS 2021 - 10th World Congress in Probability and Statistics. Seoul / Virtual, South Korea, 19th July 2021. URL: <https://hal.archives-ouvertes.fr/hal-03301609>.
- [41] M. Allouche, S. Girard and E. Gobet. ‘On the approximation of extreme quantiles with ReLU neural networks’. In: EVA 2021 - 12th International Conference on Extreme Value Analysis. Edinburgh / Virtual, United Kingdom, 28th June 2021. URL: <https://hal.archives-ouvertes.fr/hal-03301431>.

- [42] J. Arbel, M. Beraha and D. Bystrova. ‘Bayesian block-diagonal graphical models via the Fiedler prior’. In: SFdS - 52 Journées de Statistique de la Société Française de Statistique. Nice, France, 7th June 2021, pp. 1–6. URL: <https://hal.inria.fr/hal-03275245>.
- [43] J. Arbel, F. Forbes, H. D. Nguyen and T. Nguyen. ‘Approximate Bayesian computation with surrogate posteriors’. In: ISBA 2021 - World Meeting of the International Society for Bayesian Analysis. Marseille, France, 28th June 2021. URL: <https://hal.inria.fr/hal-03337949>.
- [44] J. Arbel, S. Girard, T. Moins, A. Dutfoy and K. Leachouri. ‘Improving MCMC convergence diagnostic with a local version of R-hat’. In: MAS 2021 - Journées Modélisation Aléatoire et Statistique. Orléans, France, 25th Aug. 2021. URL: <https://hal.inria.fr/hal-03337454>.
- [45] M. Bousebata, G. Enjolras and S. Girard. ‘Extreme partial least-squares regression’. In: EVA 2021 - 12th International Conference on Extreme Value Analysis. Edinburgh / Virtual, United Kingdom, 28th June 2021. URL: <https://hal.archives-ouvertes.fr/hal-03301445>.
- [46] M. Bousebata, G. Enjolras and S. Girard. ‘Extreme partial least-squares regression’. In: CMStatistics 2021 - 14th International Conference of the ERCIM WG on Computational and Methodological Statistics. London, United Kingdom, 18th Dec. 2021. URL: <https://hal.archives-ouvertes.fr/hal-03501787>.
- [47] T. Coudert, S. Ancelet, N. Pyatigorskaya, L. Nichelli, D. Ricard, D. Psimaras, M. O. Bernier, M. Dojat, F. Forbes and A. Tucholka. ‘Apport du Transfer Learning pour la segmentation automatique de lésions cérébrales radio-induites chez des patients atteints de glioblastome à partir d’un nombre restreint d’IRMs annotées’. In: Journées de biostatistique 2021 du GDR « Statistiques & Santé ». virtuel, France, 2021. URL: <https://hal.archives-ouvertes.fr/hal-03505861>.
- [48] J. El Methni and S. Girard. ‘A bias-reduced version of the Weissman estimator for extreme value-at-risk’. In: CMStatistics 2021 - 14th International Conference of the ERCIM WG on Computational and Methodological Statistics. London, United Kingdom, 18th Dec. 2021. URL: <https://hal.archives-ouvertes.fr/hal-03501777>.
- [49] J. El Methni and S. Girard. ‘A bias-reduced version of the Weissman extreme quantile estimator’. In: EVA 2021 - 12th International Conference on Extreme Value Analysis. Edinburgh / Virtual, United Kingdom, 28th June 2021. URL: <https://hal.archives-ouvertes.fr/hal-03301376>.
- [50] S. Girard and E. Gobet. ‘Estimation of the largest tail-index and extreme quantiles from a mixture of heavy-tailed distributions’. In: CMStatistics 2021 - 14th International Conference of the ERCIM WG on Computational and Methodological Statistics. London, United Kingdom, 18th Dec. 2021. URL: <https://hal.archives-ouvertes.fr/hal-03501785>.
- [51] S. Girard and E. Gobet. ‘Estimation of the tail-index and extreme quantiles from a mixture of heavy-tailed distributions’. In: RESIM 2021 - 13th International Workshop on Rare-Event Simulation. Paris / Virtual, France, 18th May 2021, p. 1. URL: <https://hal.inria.fr/hal-03235031>.
- [52] 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>.
- [53] B. Kugler, F. Forbes, S. Douté and M. Gay. ‘Efficient Bayesian data assimilation via inverse regression’. In: SFdS 2020 - 52èmes Journées de Statistiques de la Société Française de Statistique. Nice, France, 7th June 2021, pp. 1–6. URL: <https://hal.archives-ouvertes.fr/hal-03215200>.
- [54] B. Lambert, F. Forbes, S. Doyle, A. Tucholka and M. Dojat. ‘Fast Uncertainty Quantification for Deep Learning-based MR Brain Segmentation’. In: EGC 2022 - Conférence francophone pour l’Extraction et la Gestion des Connaissances. Blois, France, 24th Jan. 2022, pp. 1–12. URL: <https://hal.archives-ouvertes.fr/hal-03498120>.
- [55] B. Lambert, F. Forbes, A. Tucholka, S. Doyle and M. Dojat. ‘Multi-Scale Evaluation of Uncertainty Quantification Techniques for Deep Learning based MRI Segmentation’. In: ISMRM-ESMRMB & ISMRT 2022 - 31st Joint Annual Meeting International Society for Magnetic Resonance in Medicine. London, United Kingdom, 7th May 2022, pp. 1–3. URL: <https://hal.archives-ouvertes.fr/hal-03578023>.

- [56] T. Moins, J. Arbel, A. Dutfoy and S. Girard. ‘A Bayesian Framework for Poisson Process Characterization of Extremes with Objective Prior’. In: ISBA 2021 - World Meeting of the International Society for Bayesian Analysis. Virtual, France, 28th June 2021. URL: <https://hal.inria.fr/hal-03347871>.
- [57] T. Moins, J. Arbel, A. Dutfoy and S. Girard. ‘A Bayesian framework for Poisson process characterization of extremes with uninformative prior’. In: CMStatistics 2021 - 14th International Conference of the ERCIM WG on Computational and Methodological Statistics. London, United Kingdom, 18th Dec. 2021. URL: <https://hal.archives-ouvertes.fr/hal-03501794>.
- [58] A. Usseglio-Carleve, S. Girard and G. Stupfler. ‘Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models’. In: CMStatistics 2021 - 14th International Conference of the ERCIM WG on Computational and Methodological Statistics. London, United Kingdom, 18th Dec. 2021. URL: <https://hal.archives-ouvertes.fr/hal-03501801>.
- [59] A. Usseglio-Carleve, S. Girard and G. Stupfler. ‘Extreme expectile regression: theory and applications’. In: EVA 2021 - 12th International Conference on Extreme Value Analysis. Edinburgh / Virtual, United Kingdom, 28th June 2021. URL: <https://hal.archives-ouvertes.fr/hal-03301456>.
- [60] M. Vladimirova, J. Arbel and S. Girard. ‘Bayesian neural network unit priors and generalized Weibull-tail property’. In: ACML 2021 - 13th Asian Conference on Machine Learning. Virtual, Unknown Region, 17th Nov. 2021, pp. 1–16. URL: <https://hal.archives-ouvertes.fr/hal-03368522>.
- [61] M. Vladimirova, J. Arbel and S. Girard. ‘Generalized Weibull-tail distributions’. In: JDS 2021 - 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-03264446>.

#### Scientific book chapters

- [62] S. Girard, G. Stupfler and A. Usseglio-Carleve. ‘Extreme Lp-quantile kernel regression’. In: *Advances in Contemporary Statistics and Econometrics*. Springer, 2021, pp. 197–219. DOI: [10.1007/978-3-030-73249-3\\_11](https://doi.org/10.1007/978-3-030-73249-3_11). URL: <https://hal.archives-ouvertes.fr/hal-03182032>.

#### Doctoral dissertations and habilitation theses

- [63] B. Kugler. ‘Massive hyperspectral images analysis by inverse regression of physical models’. Université Grenoble Alpes [2020-....], 7th July 2021. URL: <https://tel.archives-ouvertes.fr/tel-03417184>.

#### Reports & preprints

- [64] S. Achard and I. Gannaz. *Whittle estimation with (quasi-)analytic wavelets*. 28th June 2021. URL: <https://hal.archives-ouvertes.fr/hal-03272326>.
- [65] M. Allouche, J. El Methni and S. Girard. *A refined Weissman estimator for extreme quantiles*. 15th Mar. 2022. URL: <https://hal.inria.fr/hal-03266676>.
- [66] M. Allouche, S. Girard and E. Gobet. *A generative model for fBm with deep ReLU neural networks*. 26th Jan. 2022. URL: <https://hal.archives-ouvertes.fr/hal-03237854>.
- [67] M. Allouche, S. Girard and E. Gobet. *EV-GAN: Simulation of extreme events with ReLU neural networks*. 16th June 2021. URL: <https://hal.archives-ouvertes.fr/hal-03250663>.
- [68] J. Arbel, S. Girard, H. D. Nguyen and A. Usseglio-Carleve. *Multivariate expectile-based distribution: properties, Bayesian inference, and applications*. 15th Nov. 2021. URL: <https://hal.inria.fr/hal-03428827>.
- [69] 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>.
- [70] M. Bousebata, G. Enjolras and S. Girard. *Extreme Partial Least-Squares regression*. 2021. URL: <https://hal.inria.fr/hal-03165399>.



- [71] 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>.
- [72] 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>.
- [73] A. Constantin, M. Fauvel and S. Girard. *Mixture of multivariate gaussian processes for classification of irregularly sampled satellite image time-series*. 2021. URL: <https://hal.archives-ouvertes.fr/hal-03280484>.
- [74] P. A. Dkengne Sielenou and S. Girard. *Estimation of upper bounds for return levels: Methodology and illustrations on univariate random sequences*. Inria - Research Centre Grenoble – Rhône-Alpes, 9th Feb. 2022. URL: <https://hal.inria.fr/hal-03562999>.
- [75] 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>.
- [76] 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>.
- [77] 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>.
- [78] M. N. Hounkonnou and P. D. Sielenou. *Conservation laws for under determined systems of differential equations*. 4th Feb. 2022. URL: <https://hal.inria.fr/hal-03557495>.
- [79] N. Mahouton and P. D. Sielenou. *Extremum conditions for functionals involving higher derivatives of several variable vector valued functions*. 4th Feb. 2022. URL: <https://hal.inria.fr/hal-03557517>.
- [80] 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>.
- [81] T. Moins, J. Arbel, A. Dutfoy and S. Girard. *On the use of a local  $\hat{R}$  to improve MCMC convergence diagnostic*. 7th Mar. 2022. URL: <https://hal.inria.fr/hal-03600407>.
- [82] H. D. Nguyen, F. Forbes, G. Fort and O. Cappé. *An online Minorization-Maximization algorithm*. 25th Jan. 2022. URL: <https://hal.archives-ouvertes.fr/hal-03542180>.
- [83] T. Nguyen, F. Chamroukhi, H. D. Nguyen and F. Forbes. *A non-asymptotic model selection in block-diagonal mixture of polynomial experts models*. 2nd May 2021. URL: <https://hal.archives-ouvertes.fr/hal-03214604>.
- [84] T. Nguyen, H. D. Nguyen, F. Chamroukhi and F. Forbes. *A non-asymptotic penalization criterion for model selection in mixture of experts models*. 2nd May 2021. URL: <https://hal.archives-ouvertes.fr/hal-03214606>.
- [85] B. Olivier, A. Guérin-Dugué and J.-B. Durand. *Hidden Semi-Markov Models to Segment Reading Phases from Eye Movements*. RR-9398. Inria Grenoble - Rhône-Alpes, 2nd Mar. 2021. URL: <https://hal.inria.fr/hal-03155843>.

### Other scientific publications

- [86] A. Delphin, F. Boux, C. Brossard, J. M. Warnking, B. Lemasson, E. L. Barbier and T. Christen. ‘Optimisation des patterns de signaux pour l’IRM Fingerprinting Vasculaire’. In: SFRMBM 2021 - 5ème Congrès scientifique de la Société Française de Résonance Magnétique en Biologie et Médecine. Lyon, France, 27th Sept. 2021, pp. 1–1. URL: <https://hal.univ-grenoble-alpes.fr/hal-03418716>.
- [87] T. Moins, J. Arbel, S. Girard and A. Dutfoy. ‘Improving MCMC convergence diagnostic: a local version of R-hat’. In: BayesComp-ISBA workshop: Measuring the quality of MCMC output. online, France, 6th Oct. 2021. URL: <https://hal.inria.fr/hal-03372736>.

- [88] S. Potin Manigand, S. Douté, B. Kugler and F. Forbes. ‘Comparison of photometric phase curves resulting from various observation scenes’. In: EPSC 2021 - Europlanet Science Congress. Grenoble, France, 13th Sept. 2021, pp. 1–1. URL: <https://hal.archives-ouvertes.fr/hal-03498138>.

### 11.3 Other

#### Scientific popularization

- [89] J.-B. Durand and S. Achard. ‘Comprendre un processus cognitif grâce à l’analyse statistique du mouvement des yeux’. In: *Interstices* (17th Dec. 2021). URL: <https://hal.inria.fr/hal-03505113>.

### 11.4 Cited publications

- [90] 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>.
- [91] P. Embrechts, C. Klüppelberg and T. Mikosh. *Modelling Extremal Events*. Vol. 33. Applications of Mathematics. Springer-Verlag, 1997.
- [92] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics, Springer, 2006.
- [93] 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 Cergy-Pontoise, Oct. 1996.
- [94] K. Li. ‘Sliced inverse regression for dimension reduction’. In: *Journal of the American Statistical Association* 86 (1991), pp. 316–327.
- [95] 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.