IN PARTNERSHIP WITH:
**Institut polytechnique de Grenoble**

**Université de Grenoble Alpes**

Activity Report 2019

# Project-Team MISTIS

## Modelling and Inference of Complex and Structured Stochastic Systems

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
**Grenoble - Rhône-Alpes**

THEME
**Optimization, machine learning and statistical methods**

# Table of contents

# Project-Team MISTIS

*Creation of the Project-Team: 2008 January 01*

**Keywords:**

**Computer Science and Digital Science:**

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, Visitors, External Collaborators

**Research Scientists**

Florence Forbes [Team leader, Inria, Senior Researcher, HDR]
Sophie Achard [CNRS, Senior Researcher, from Sep 2019, HDR]
Julyan Arbel [Inria, Researcher, HDR]

Stephane Girard [Inria, Senior Researcher, HDR]

**Faculty Member**

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

**Post-Doctoral Fellows**

Alexis Arnaud [Inria, Post-Doctoral Fellow, until Jul 2019]
Marta Crispino [Inria, Post-Doctoral Fellow, until Mar 2019]
Pascal Dkengne Sielenou [Inria, Post-Doctoral Fellow, until Oct 2019]
Hongliang Lu [Inria, Post-Doctoral Fellow, until Jan 2019]
Antoine Usseglio Carleve [Inria, Post-Doctoral Fellow]
Fei Zheng [Inria, Post-Doctoral Fellow]

**PhD Students**

Karina Ashurbekova [Univ Grenoble Alpes, PhD Student]
Meryem Bousebata [Univ Grenoble Alpes, PhD Student]
Fabien Boux [Univ Grenoble Alpes, PhD Student]
Daria Bystrova [Univ Grenoble Alpes, PhD Student, from Oct 2019]
Alexandre Constantin [Univ Grenoble Alpes, PhD Student]
Benoit Kugler [Univ Grenoble Alpes, PhD Student]
Veronica Munoz Ramirez [Univ Grenoble Alpes, PhD Student]
Brice Olivier [Univ Grenoble Alpes, PhD Student, until May 2019]
Giovanni Poggiato [Univ Grenoble Alpes, PhD Student, from Nov 2019]
Mariia Vladimirova [Inria, PhD Student]

**Technical staff**

Fatima Fofana [Inria, Engineer, until Aug 2019]

**Interns and Apprentices**

Daria Bystrova [Institut polytechnique de Grenoble, from Mar 2019 until Jul 2019]
Valentin Chevalier [Inria, from Apr 2019 until Sep 2019]
Fatoumata Dama [ENSIMAG, from Mar 2019 until Jul 2019]
Virgilio Kmetzsch Rosa E Silva [Inria, from Feb 2019 until Jul 2019]
Sharan Yalburgi [Inria, from Jun 2019 until Jul 2019]

**Administrative Assistant**

Marion Ponsot [Inria, Administrative Assistant]

**Visiting Scientists**

Aboubacrène Ag Ahmad [Univ. Gaston Berger, Senegal, from Nov 2019]
Hien Nguyen [Inria, from Nov 2019]
Darren Wraith [QUT, Brisbane, Australia, from Dec 2019]

**External Collaborators**

Sophie Achard [CNRS, from May 2019 until Aug 2019, HDR]
Thibaud Rahier [Criteo, from Feb 2019 until Apr 2019]

# 2. Overall Objectives

## 2.1. Overall Objectives

The context of our work is the analysis of structured stochastic models with statistical tools. The idea underlying the concept of structure is that stochastic systems that exhibit great complexity can be accounted for by combining simple local assumptions in a coherent way. This provides a key to modelling, computation, inference and interpretation. This approach appears to be useful in a number of high impact applications including signal and image processing, neuroscience, genomics, sensors networks, etc. while the needs from these domains can in turn generate interesting theoretical developments. However, this powerful and

flexible approach can still be restricted by necessary simplifying assumptions and several generic sources of complexity in data.

Often data exhibit complex dependence structures, having to do for example with repeated measurements on individual items, or natural grouping of individual observations due to the method of sampling, spatial or temporal association, family relationship, and so on. Other sources of complexity are related to the measurement process, such as having multiple measuring instruments or simulations generating high dimensional and heterogeneous data or such that data are dropped out or missing. Such complications in data-generating processes raise a number of challenges. Our goal is to contribute to statistical modelling by offering theoretical concepts and computational tools to handle properly some of these issues that are frequent in modern data. So doing, we aim at developing innovative techniques for high scientific, societal, economic impact applications and in particular via image processing and spatial data analysis in environment, biology and medicine.

The methods we focus on involve mixture models, Markov models, and more generally hidden structure models identified by stochastic algorithms on one hand, and semi and non-parametric methods on the other hand.

Hidden structure models are useful for taking into account heterogeneity in data. They concern many areas of statistics (finite mixture analysis, hidden Markov models, graphical models, random effect models, ...). Due to their missing data structure, they induce specific difficulties for both estimating the model parameters and assessing performance. The team focuses on research regarding both aspects. We design specific algorithms for estimating the parameters of missing structure models and we propose and study specific criteria for choosing the most relevant missing structure models in several contexts.

Semi and non-parametric methods are relevant and useful when no appropriate parametric model exists for the data under study either because of data complexity, or because information is missing. When observations are curves, they enable us to model the data without a discretization step. These techniques are also of great use for *dimension reduction* purposes. They enable dimension reduction of the functional or multivariate data with no 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 which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis*, which is based on the modelling of distribution tails by both a functional part and a real parameter.

# 3. Research Program

## 3.1. Mixture models

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

**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 multidimensional, usually unknown and to be estimated. We consider cases where the data naturally divides into observed data $y = \{y_1, ..., y_n\}$ and unobserved or missing data $z = \{z_1, ..., z_n\}$. The missing data $z_i$ represents for instance the memberships of one of a set of $K$ alternative categories. The distribution of an observed $y_i$ can be written as a finite mixture of distributions,

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

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

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

## 3.2. Markov models

**Participants:** Alexis Arnaud, Brice Olivier, Jean-Baptiste Durand, Florence Forbes, Karina Ashurbekova, Hongliang Lu, Julyan Arbel, Mariia Vladimirova.

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

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

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

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

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

## 3.3. Functional Inference, semi- and non-parametric methods

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

**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 [82]. 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 [84] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [81], 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 $X_{1,n} \leq ... \leq X_{n,n}$ denote $n$ ordered observations from a random variable $X$ representing some quantity of interest. A $p_n$-quantile of $X$ is the value $x_{p_n}$ such that the probability that $X$ is greater than $x_{p_n}$ is $p_n$, *i.e.* $P(X > x_{p_n}) = p_n$. When $p_n < 1/n$, such a quantile is said to be extreme since it is usually greater than the maximum observation $X_{n,n}$.

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

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

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

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

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

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

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

### *3.3.2. 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 a possible loss of information. Principal Component Analysis (PCA) is the most widely used method to reduce dimension in data. However, standard linear PCA can be quite inefficient on image data where even simple image distortions can lead to highly non-linear data. Two directions are investigated. First, non-linear PCAs can be proposed, leading to semi-parametric dimension reduction methods [83]. 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 [80]. 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 [84].

# 4. Application Domains

## 4.1. Image Analysis

**Participants:** Alexis Arnaud, Veronica Munoz Ramirez, Florence Forbes, Stephane Girard, Hongliang Lu, Fabien Boux, Benoit Kugler, Alexandre Constantin.

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

## 4.2. Biology, Environment and Medicine

**Participants:** Alexis Arnaud, Florence Forbes, Stephane Girard, Jean-Baptiste Durand, Julyan Arbel, Brice Olivier, Karina Ashurbekova, Fabien Boux, Veronica Munoz Ramirez, Fei Zheng.

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

# 5. Highlights of the Year

## 5.1. Highlights of the Year

**New appointments:**

- Florence Forbes has been appointed as a member of the advisory committee of the Helmholtz AI Cooperation Unit https://helmholtz.ai/.

**Data Challenges**

- Pixyl winner of the Société Francaise de Radiologie Data Challenge 2019

  Pixyl, a Grenoble-based start-up originating in the team and Inserm, accompanied by a team of neuroradiologists and academics, distinguished itself in the AI challenge held during the 2019 edition of the Journées Francophone de Radiologie, which took place from 11 to 14 October in Paris. The Challenge was about prediction of multiple sclerosis patient disability from a single MRI image

### *5.1.1. Awards*

- Meryem Bousebata received the second best presentation award at the "10th conference of the international society for Integrated Disaster Risk Management (IDRiM)" organized by CNRS-University of Nice and AFPCN and held from 16 to 18 October 2019 in Nice.
- Mariia Vladimirova received the best poster award for her work [45] at the "12th Conference on Bayesian Nonparametrics", Oxford University, UK, June 24-28, 2019.

BEST PAPER AWARD:

[52]
M. BOUSEBATA, G. ENJOLRAS, S. GIRARD. *Bayesian estimation of natural extreme risk measures. Application to agricultural insurance*, in "IDRiM 2019 - 10th conference of the international society for Integrated Disaster Risk Management", Nice, France, October 2019, https://hal.archives-ouvertes.fr/hal-02276292

# 6. New Software and Platforms

## 6.1. BOLD model FIT

KEYWORDS: Functional imaging - FMRI - Health

SCIENTIFIC DESCRIPTION: Physiological and biophysical models have been proposed to link neuronal activity to the Blood Oxygen Level-Dependent (BOLD) signal in functional MRI (fMRI). Those models rely on a set of parameter values that are commonly estimated using gradient-based local search methods whose initial values are taken from the literature. In some applications, interesting insight into the brain physiology or physiopathology can be gained from an estimation of the model parameters from measured BOLD signals. In this work we focus on the extended Balloon model and propose the estimation of 15 parameters using seven different approaches: three versions of the Expectation Maximization Gauss-Newton (EM/GN) approach (the *de facto* standard in the neuroscientific community) and four metaheuristics (Particle Swarm Optimization (PSO), Differential Evolution (DE), Real-Coded Genetic Algorithms (GA), and a Memetic Algorithm (MA) combining EM/GN and DE). To combine both the ability to escape local optima and to incorporate prior knowledge, we derive the target function from Bayesian modeling. The general behavior of these algorithms is analyzed and compared, providing very promising results on challenging real and synthetic fMRI data sets involving rats with epileptic activity. These stochastic optimizers provided a better performance than EM/GN in terms of distance to the ground truth in 4 out of 6 synthetic data sets and a better signal fitting in 12 out of 12 real data sets. Non-parametric statistical tests showed the existence of statistically significant differences between the real data results obtained by DE and EM/GN. Finally, the estimates obtained from DE for these parameters seem both more realistic and more stable or at least as stable across sessions as the estimates from EM/GN. This is the largest comparison of optimizers for the estimation of biophysical parameters in BOLD fMRI

FUNCTIONAL DESCRIPTION: This Matlab toolbox performs the automatic estimation of biophysical parameters using the extended Balloon model and BOLD fMRI data. It takes as input a MAT file and provides as output the parameter estimates achieved by using stochastic optimization

NEWS OF THE YEAR: The main differences with our previous work: 1) we also use synthetic data, 2) we use stochastic GN and MCMC+DE, 3) We evaluate results not only in physiological terms but also comparing fitness function values. Also changes were made to allow running on the cluster via MPI

- Participants: Pablo Mesejo Santiago, Florence Forbes and Jan Warnking
- Partner: University of Granada, Spain
- Contact: Pablo Mesejo Santiago
- Publication: A differential evolution-based approach for fitting a nonlinear biophysical model to fMRI BOLD data
- URL: https://hal.archives-ouvertes.fr/hal-01221115v2/

## 6.2. PyHRF

KEYWORDS: Medical imaging - Health - Brain - IRM - Neurosciences - Statistic analysis - FMRI

SCIENTIFIC DESCRIPTION: Functional Magnetic Resonance Imaging (fMRI) is a neuroimaging technique that allows the non-invasive study of brain function. It is based on the hemodynamic variations induced by changes in cerebral synaptic activity following sensory or cognitive stimulation. The measured signal depends on the variation of blood oxygenation level (BOLD signal) which is related to brain activity: a decrease in deoxyhemoglobin concentration induces an increase in BOLD signal. The BOLD signal is delayed with respect to changes in synaptic activity, which can be modeled as a convolution with the Hemodynamic Response Function (HRF) whose exact form is unknown and fluctuates with various parameters such as age, brain region or physiological conditions. In this work we propose to analyze fMRI data using a Joint Detection-Estimation (JDE) approach. It jointly detects cortical activation and estimates the HRF. In contrast to existing tools, PyHRF estimates the HRF instead of considering it as a given constant in the entire brain.

FUNCTIONAL DESCRIPTION: As part of fMRI data analysis, PyHRF provides a set of tools for addressing the two main issues involved in intra-subject fMRI data analysis : (i) the localization of cerebral regions that elicit evoked activity and (ii) the estimation of the activation dynamics also referenced to as the recovery of the Hemodynamic Response Function (HRF). To tackle these two problems, PyHRF implements the Joint Detection-Estimation framework (JDE) which recovers parcel-level HRFs and embeds an adaptive spatio-temporal regularization scheme of activation maps.

NEWS OF THE YEAR: The framework to perform software tests has been further developed. Some unitary tests have been set.

- Participants: Aina Frau Pascual, Christine Bakhous, Florence Forbes, Jaime Eduardo Arias Almeida, Laurent Risser, Lotfi Chaari, Philippe Ciuciu, Solveig Badillo, Thomas Perret and Thomas Vincent
- Partners: CEA - NeuroSpin
- Contact: Florence Forbes
- Publications: Flexible multivariate hemodynamics fMRI data analyses and simulations with PyHRF - Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach - A Bayesian Non-Parametric Hidden Markov Random Model for Hemodynamic Brain Parcellation
- URL: http://pyhrf.org

## 6.3. xLLiM

*High dimensional locally linear mapping*

KEYWORDS: Clustering - Regression

SCIENTIFIC DESCRIPTION: Building a regression model for the purpose of prediction is widely used in all disciplines. A large number of applications consists of learning the association between responses and predictors and focusing on predicting responses for the newly observed samples. In this work, we go beyond simple linear models and focus on predicting low-dimensional responses using high-dimensional covariates when the associations between responses and covariates are non-linear.

FUNCTIONAL DESCRIPTION: This is an R package available on the CRAN at https://cran.r-project.org/web/packages/xLLiM/index.html

XLLiM provides a tool for non linear mapping (non linear regression) using a mixture of regression model and an inverse regression strategy. The methods include the GLLiM model (Deleforge et al (2015) ) based on Gaussian mixtures and a robust version of GLLiM, named SLLiM (see Perthame et al (2016) ) based on a mixture of Generalized Student distributions.

NEWS OF THE YEAR: A new Hierarchical version of GLLiM has been developed in collaboration with University of Michigan, USA.

- Participants: Antoine Deleforge, Emeline Perthame and Florence Forbes
- Partner: University of Michigan, Ann Arbor, USA
- Contact: Florence Forbes
- Publications: Inverse regression approach to robust nonlinear high-to-low dimensional mapping - High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables
- URL: https://cran.r-project.org/web/packages/xLLiM/index.html

## 6.4. MMST

*Mixtures of Multiple Scaled Student T distributions*

KEYWORDS: Health - Statistics - Brain MRI - Medical imaging - Robust clustering

SCIENTIFIC DESCRIPTION: A new family of multivariate heavy-tailed distributions that allow variable marginal amounts of tailweight is proposed and implemented. The originality comes from introducing multidimensional instead of univariate scale variables for the mixture of scaled Gaussian family of distributions. In contrast to most existing approaches, the derived distributions can account for a variety of shapes and have a simple tractable form with a closed-form probability density function whatever the dimension. We provide maximum likelihood estimation of the parameters and illustrate their modelling flexibility.

FUNCTIONAL DESCRIPTION: The package implements mixtures of so-called multiple scaled Student distributions, which are generalisation of multivariate Student T distribution allowing different tails in each dimension. Typical applications include Robust clustering to analyse data with possible outliers. In this context, the model and package have been used on large data sets of brain MRI to segment and identify brain tumors. Recent additions include a Markov random field implementation to account for spatial dependencies between observations, and a Bayesian implementation that can be used to select the number of mixture components automatically.

RELEASE FUNCTIONAL DESCRIPTION: Recent additions include a Markov random field implementation to account for spatial dependencies between observations, and a Bayesian implementation that can be used to select the number of mixture components automatically.

NEWS OF THE YEAR: Recent additions include a Markov random field implementation to account for spatial dependencies between observations, and a Bayesian implementation that can be used to select the number of mixture components automatically.

- Participants: Alexis Arnaud, Darren Wraith, Florence Forbes, Steven Quinito Masnada and Stéphane Despréaux
- Partner: Institut des Neurosciences Grenoble
- Contact: Florence Forbes
- Publications: A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering - Fully Automatic Lesion Localization and Characterization: Application to Brain Tumors Using Multiparametric Quantitative MRI Data
- URL: https://team.inria.fr/mistis/software/

# 7. New Results

## 7.1. Mixture models

### *7.1.1. Mini-batch learning of exponential family finite mixture models*

**Participant:** Florence Forbes.

**Joint work with**: Hien Nguyen, La Trobe University Melbourne Australia and Geoffrey J. McLachlan, University of Queensland, Brisbane, Australia.

Mini-batch algorithms have become increasingly popular due to the requirement for solving optimization problems, based on large-scale data sets. Using an existing online expectation-maximization (EM) algorithm framework, we demonstrate [28] how mini-batch (MB) algorithms may be constructed, and propose a scheme for the stochastic stabilization of the constructed mini-batch algorithms. Theoretical results regarding the convergence of the mini-batch EM algorithms are presented. We then demonstrate how the mini-batch framework may be applied to conduct maximum likelihood (ML) estimation of mixtures of exponential family distributions, with emphasis on ML estimation for mixtures of normal distributions. Via a simulation study, we demonstrate that the mini-batch algorithm for mixtures of normal distributions can outperform the standard EM algorithm. Further evidence of the performance of the mini-batch framework is provided via an application to the famous MNIST data set.

### *7.1.2. Component elimination strategies to fit mixtures of multiple scale distributions*

**Participants:** Florence Forbes, Alexis Arnaud.

We address the issue of selecting automatically the number of components in mixture models with non-Gaussian components. As amore efficient alternative to the traditional comparison of several model scores in a range, we consider procedures based on a single run of the inference scheme. Starting from an over-fitting mixture in a Bayesian set-ting, we investigate two strategies to eliminate superfluous components.We implement these strategies for mixtures of multiple scale distributions which exhibit a variety of shapes not necessarily elliptical while remaining analytical and tractable in multiple dimensions. A Bayesian formulation and a tractable inference procedure based on variational approximation are proposed. Preliminary results on simulated and real data show promising performance in terms of model selection and computational time. This work has been presented at RSSDS 2019 - Research School on Statistics and Data Science in Melbourne, Australia [33].

### *7.1.3. Approximate Bayesian Inversion for high dimensional problems*

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

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

The overall objective is to develop a statistical learning technique capable of solving complex inverse problems in setting with specific constraints. More specifically, the challenges are 1) the large number of observations to be inverted, 2) their large dimension, 3) the need to provide predictions for correlated parameters and 4) the need to provide a quality index (eg. uncertainty).

In the context of Bayesian inversion, one can use a regression approach, such as in the so-called Gaussian Locally Linear Mapping (GLLiM) [7], to obtain an approximation of the posterior distribution. In some cases, exploiting this approximate distribution remains challenging, for example because of its multi-modality. In this work, we investigate the possible use of Importance Sampling to build on the standard GLLiM approach by improving the approximation induced by the method and to better handle the potential existence of multiple solutions. We may also consider our approach as a way to provide an informed proposal distribution as requested by Importance Sampling techniques. We experiment our approach on simulated and real data in the context of a photometric model inversion in planetology. Preliminary results have been presented at StatLearn 2019 [76]

### *7.1.4. MR fingerprinting parameter estimation via inverse regression*

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

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

Magnetic resonance imaging (MRI) can map a wide range of tissue properties but is often limited to observe a single parameter at a time. In order to overcome this problem, Ma et al. introduced magnetic resonance fingerprinting (MRF), a procedure based on a dictionary of simulated couples of signals and parameters. Acquired signals called fingerprints are then matched to the closest signal in the dictionary in order to estimate parameters. This requires an exhaustive search in the dictionary, which even for moderately sized problems, becomes costly and possibly intractable . We propose an alternative approach to estimate more parameters at a time. Instead of an exhaustive search for every signal, we use the dictionary to learn the functional relationship between signals and parameters. . A dictionary-based learning (DBL) method was investigated to bypass inherent MRF limitations in high dimension: reconstruction time and memory requirement. The DBL method is a 3-step procedure: (1) a quasi-random sampling strategy to produce the dictionary, (2) a statistical inverse regression model to learn from the dictionary a probabilistic mapping between MR fingerprints and parameters, and (3) this mapping to provide both parameter estimates and their confidence levels. On synthetic data, experiments show that the quasi-random sampling outperforms the grid when designing the dictionary for inverse regression. Dictionaries up to 100 times smaller than usually employed in MRF yield more accurate parameter estimates with a 500 time gain. Estimates are supplied with a confidence index, well correlated with the estimation bias. On microvascular MRI data, results showed that dictionary-based methods (MRF and DBL) yield more accurate estimates than the conventional, closed-form equation, method. On MRI signals from tumor bearing rats, the DBL method shows very little sensitivity to the dictionary size in contrast to the

MRF method. The proposed method efficiently reduces the number of required simulations to produce the dictionary, speeds up parameter estimation, and improve estimates accuracy. The DBL method also introduces a confidence index for each parameter estimate. Preliminary results have been presented at the third *Congrés National d'Imagerie du Vivant* (CNIV 2019) [53] and at the fourth *Congrés de la Société Française de Résonance Magnétique en Biologie et Médecine* (SFRMBM 2019) [54].

### *7.1.5. Characterization of daily glycemic variability in subjects with type 1 diabetes using a mixture of metrics*

**Participants:** Florence Forbes, Fei Zheng.

**Joint work with**: Stéphane Bonnet from CEA Leti and Pierre-Yves Benhamou, Manon Jalbert from CHU Grenoble Alpes.

Glycemic variability is an important component of glycemic control for patients with type 1 diabetes. Glycemic variability (GV) must be taken into account in the efficacy of treatment of type 1 diabetes because it determines the quality of glycemic control, the risk of complication of the patient's disease. In a first study [24], our goal was to describe GV scores in patients with pancreatic islet transplantation (PIT) type 1 diabetes in the TRIMECO trial, and change of thresholds, for each index. predictive of success of PIT.

In a second study, we address the issue of choosing an appropriate measure of GV. Many metrics have been proposed to account for this variability but none is unanimous among physicians. The inadequacy of existing measurements lies in the fact that they view the variability from different aspects, so that no consensus has been reached among physicians as to which metrics to use in practice. Moreover, although glycemic variability, from one day to another, can show very different patterns, few metrics have been dedicated to daily evaluations. In this work [50], [30], a reference (stable-glycemia) statistical model is built based on a combination of daily computed canonical glycemic control metrics including variability. The metrics are computed for subjects from the TRIMECO islet transplantation trial , selected when their $\beta$-score (composite score for grading success) is greater than 6 after a transplantation. Then, for any new daily glycemia recording, its likelihood with respect to this reference model provides a multi-metric score of daily glycemic variability severity. In addition, determining the likelihood value that best separates the daily glycemia with a zero $\beta$-score from that greater than 6, we propose an objective decision rule to classify daily glycemia into "stable" or "unstable". The proposed characterization framework integrates multiple standard metrics and provides a comprehensive daily glycemic variability index, based on which, long term variability evaluations and investigations on the implicit link between variability and $\beta$-score can be carried out. Evaluation, in a daily glycemic variability classification task, shows that the proposed method is highly concordant to the experience of diabetologists. A multivariate statistical model is therefore proposed to characterize the daily glycemic variability of subjects with type 1 diabetes. The model has the advantage to provide a single variability score that gathers the information power of a number of canonical scores, too partial to be used individually. A reliable decision rule to classify daily variability measurements into stable or unstable is also provided.

### *7.1.6. Dirichlet process mixtures under affine transformations of the data*

**Participant:** 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 [63], 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.

### *7.1.7. Approximate Bayesian computation via the energy statistic*

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

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

Approximate Bayesian computation (ABC) has become an essential part of the Bayesian toolbox for addressing problems in which the likelihood is prohibitively expensive or entirely unknown, making it intractable. ABC defines a quasi-posterior by comparing observed data with simulated data, traditionally based on some summary statistics, the elicitation of which is regarded as a key difficulty. In recent years, a number of data discrepancy measures bypassing the construction of summary statistics have been proposed, including the Kullback-Leibler divergence, the Wasserstein distance and maximum mean discrepancies. In this work [79], we propose a novel importance-sampling (IS) ABC algorithm relying on the so-called two-sample energy statistic. We establish a new asymptotic result for the case where both the observed sample size and the simulated data sample size increase to infinity, which highlights to what extent the data discrepancy measure impacts the asymptotic pseudo-posterior. The result holds in the broad setting of IS-ABC methodologies, thus generalizing previous results that have been established only for rejection ABC algorithms. Furthermore, we propose a consistent V-statistic estimator of the energy statistic, under which we show that the large sample result holds. Our proposed energy statistic based ABC algorithm is demonstrated on a variety of models, including a Gaussian mixture, a moving-average model of order two, a bivariate beta and a multivariate g-and-k distribution. We find that our proposed method compares well with alternative discrepancy measures.

### *7.1.8. Industrial applications of mixture modeling*

**Participant:** Julyan Arbel.

**Joint work with**: Kerrie Mengersen and Earl Duncan from QUT, School of Mathematical Sciences, Brisbane, Australia, and Clair Alston-Knox, Griffith University Brisbane, Australia, and Nicole White, Institute for Health and Biomedical Innovation, Brisbane, Australia.

In [61], we illustrate the wide diversity of applications of mixture models to problems in industry, and the potential advantages of these approaches, through a series of case studies. The first of these focuses on the iconic and pervasive need for process monitoring, and reviews a range of mixture approaches that have been proposed to tackle complex multimodal and dynamic or online processes. The second study reports on mixture approaches to resource allocation, applied here in a spatial health context but which are applicable more generally. The next study provides a more detailed description of a multivariate Gaussian mixture approach to a biosecurity risk assessment problem, using big data in the form of satellite imagery. This is followed by a final study that again provides a detailed description of a mixture model, this time using a nonparametric formulation, for assessing an industrial impact, notably the influence of a toxic spill on soil biodiversity.

## 7.2. Semi and non-parametric methods

### *7.2.1. Deep learning models to study the early stages of Parkinson's Disease*

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

**Joint work with**: Michel Dojat from Grenoble Institute of Neuroscience.

Current physio-pathological data suggest that Parkinson's Disease (PD) symptoms are related to important alterations in subcortical brain structures. However, structural changes in these small structures remain difficult to detect for neuro-radiologists, in particular, at the early stages of the disease (*de novo* PD patients) [58], [43], [59]. The absence of a reliable ground truth at the voxel level prevents the application of traditional supervised deep learning techniques. In this work, we consider instead an anomaly detection approach and show that auto-encoders (AE) could provide an efficient anomaly scoring to discriminate *de novo* PD patients using quantitative Magnetic Resonance Imaging (MRI) data.

### *7.2.2. Estimation of extreme risk measures*

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

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

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. The Value-at-Risk however suffers from several weaknesses. First, it provides us only with a pointwise information: 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.

However, the asymptotic normality of the empirical CTE estimator requires that the underlying distribution possess a finite variance; this can be a strong restriction in heavy-tailed models which constitute the favoured class of models in actuarial and financial applications. One possible solution in very heavy-tailed models where this assumption fails could be to use the more robust Median Shortfall, but this quantity is actually just a quantile, which therefore only gives information about the frequency of a tail event and not about its typical magnitude. In [23], we construct a synthetic class of tail $L_p$-medians, which encompasses the Median Shortfall (for $p = 1$) and Conditional Tail Expectation (for $p = 2$). We show that, for $1 < p < 2$, a tail $L_p$-median always takes into account both the frequency and magnitude of tail events, and its empirical estimator is, within the range of the data, asymptotically normal under a condition weaker than a finite variance. We extrapolate this estimator, along with another technique, to proper extreme levels using the heavy-tailed framework. The estimators are showcased on a simulation study and on a set of real fire insurance data showing evidence of a very heavy right tail.

A possible coherent alternative risk measure is based on expectiles [6]. Compared to quantiles, the family of expectiles is based on squared rather than absolute error loss minimization. The flexibility and virtues of these least squares analogues of quantiles are now well established in actuarial science, econometrics and statistical finance. have recently received a lot of attention, especially in actuarial and financial risk man-agement. Their estimation, however, typically requires to consider non-explicit asymmetric leasts-quares estimates rather than the traditional order statistics used for quantile estimation. This makes the study of the tail expectile process a lot harder than that of the standard tail quantile process. Under the challenging model of heavy-tailed distributions, we derive joint weighted Gaussian approximations of the tail empirical expectile and quantile processes. We then use this powerful result to introduce and study new estimators of extreme expectiles and the standard quantile-based expected shortfall, as well as a novel expectile-based form of expected shortfall [22].

Both quantiles and expectiles were embedded in the more general class of $L_p$-quantiles [21] as the minimizers of a generic asymmetric convex loss function. It has been proved very recently that the only $L_p$-quantiles that are coherent risk measures are the expectiles. In [75], we work in a context of heavy tails, which is especially relevant to actuarial science, finance, econometrics and natural sciences, and we construct an estimator of the tail index of the underlying distribution based on extreme $L_p$-quantiles. We establish the asymptotic normality of such an estimator and in doing so, we extend very recent results on extreme expectile and $L_p$-quantile estimation. We provide a discussion of the choice of $p$ in practice, as well as a methodology for reducing the bias of our estimator. Its finite-sample performance is evaluated on simulated data and on a set of real hydrological data. This work is submitted for publication.

### *7.2.3. Conditional extremal events*

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

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

The goal of the PhD thesis of Aboubacrene Ag Ahmad is to contribute to the development of theoretical and algorithmic models to tackle conditional extreme value analysis, *ie* the situation where some covariate information $X$ is recorded simultaneously with a quantity of interest $Y$. In such a case, extreme quantiles and expectiles are functions of the covariate. In [13], we consider a location-scale model for conditional heavy-tailed distributions when the covariate is deterministic. First, nonparametric estimators of the location and scale functions are introduced. Second, an estimator of the conditional extreme-value index is derived. The asymptotic properties of the estimators are established under mild assumptions and their finite sample properties are illustrated both on simulated and real data.

As explained in Paragraph 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. In [74], 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 results are submitted for publication.

### *7.2.4. Estimation of the variability in the distribution tail*

**Participant:** 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 submitted for publication [72].

### *7.2.5. Extrapolation limits associated with extreme-value methods*

**Participant:** Stephane Girard.

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

The PhD thesis of Clément Albert (co-funded by EDF) is dedicated to the study of the sensitivity of extreme-value methods to small changes in the data and to their extrapolation ability. Two directions are explored:

(i) In [15], we investigate the asymptotic behavior of the (relative) extrapolation error associated with some estimators of extreme quantiles based on extreme-value theory. It is shown that the extrapolation error can be interpreted as the remainder of a first order Taylor expansion. Necessary and sufficient conditions are then provided such that this error tends to zero as the sample size increases. Interestingly, in case of the so-called Exponential Tail estimator, these conditions lead to a subdivision of Gumbel maximum domain of attraction into three subsets. In constrast, the extrapolation error associated with Weissman estimator has a common behavior over the whole Fréchet maximum domain of attraction. First order equivalents of the extrapolation error are then derived and their accuracy is illustrated numerically.

(ii) In [14], We propose a new estimator for extreme quantiles under the log-generalized Weibull-tail model, introduced by Cees de Valk. This model relies on a new regular variation condition which, in some situations, permits to extrapolate further into the tails than the classical assumption in extreme-value theory. The asymptotic normality of the estimator is established and its finite sample properties are illustrated both on simulated and real datasets.

### 7.2.6. Bayesian inference for copulas

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

We study in [16] a broad class of asymmetric copulas known as Liebscher copulas and defined as a combination of multiple—usually symmetric—copulas. The main thrust of this work is to provide new theoretical properties including exact tail dependence expressions and stability properties. A subclass of Liebscher copulas obtained by combining Fréchet copulas is studied in more details. We establish further dependence properties for copulas of this class and show that they are characterized by an arbitrary number of singular components. Furthermore, we introduce a novel iterative construction for general Liebscher copulas which *de facto* insures uniform margins, thus relaxing a constraint of Liebscher's original construction. Besides, we show that this iterative construction proves useful for inference by developing an Approximate Bayesian computation sampling scheme. This inferential procedure is demonstrated on simulated data.

### 7.2.7. Approximations of Bayesian nonparametric models

**Participant:** Julyan Arbel.

**Joint work with**: Stefano Favaro and Pierpaolo De Blasi from Collegio Carlo Alberto, Turin, Italy, Igor Prunster from Bocconi University, Milan, Italy, Caroline Lawless from Université Paris-Dauphine, France, Olivier Marchal from Université Jean Monnet.

For a long time, the Dirichlet process has been the gold standard discrete random measure in Bayesian nonparametrics. The Pitman–Yor process provides a simple and mathematically tractable generalization, allowing for a very flexible control of the clustering behaviour. Two commonly used representations of the Pitman–Yor process are the stick-breaking process and the Chinese restaurant process. The former is a constructive representation of the process which turns out very handy for practical implementation, while the latter describes the partition distribution induced. Obtaining one from the other is usually done indirectly with use of measure theory. In contrast, we propose in [25] an elementary proof of Pitman–Yor's Chinese Restaurant process from its stick-breaking representation.

In [17], we consider approximations to the popular Pitman-Yor process obtained by truncating the stick-breaking representation. The truncation is determined by a random stopping rule that achieves an almost sure control on the approximation error in total variation distance. We derive the asymptotic distribution of the random truncation point as the approximation error goes to zero in terms of a polynomially tilted positive stable random variable. The practical usefulness and effectiveness of this theoretical result is demonstrated by devising a sampling algorithm to approximate functionals of the-version of the Pitman–Yor process.

In [18], 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 [18], we prove a monotonicity property of the Hurwitz zeta function which, in turn, translates into a chain of inequalities for polygamma functions of different orders. We provide a probabilistic interpretation of our result by exploiting a connection between Hurwitz zeta function and the cumulants of the exponential-beta distribution.

### 7.2.8. Concentration inequalities

**Participant:** Julyan Arbel.

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

In [19], we investigate the sub-Gaussian property for almost surely bounded random variables. If sub-Gaussianity per se is de facto ensured by the bounded support of said random variables, then exciting research avenues remain open. Among these questions is how to characterize the optimal sub-Gaussian proxy variance? Another question is how to characterize strict sub-Gaussianity, defined by a proxy variance equal to the (standard) variance? We address the questions in proposing conditions based on the study of functions variations. A particular focus is given to the relationship between strict sub-Gaussianity and symmetry of the distribution. In particular, we demonstrate that symmetry is neither sufficient nor necessary for strict sub-Gaussianity. In contrast, simple necessary conditions on the one hand, and simple sufficient conditions on the other hand, for strict sub-Gaussianity are provided. These results are illustrated via various applications to a number of bounded random variables, including Bernoulli, beta, binomial, uniform, Kumaraswamy, and triangular distributions.

### *7.2.9. Extraction and data analysis toward "industry of the future"*

**Participants:** Florence Forbes, Hongliang Lu, Fatima Fofana.

**Joint work with**: J. F. Cuccaro and J. C Trochet from Vi-Technology company.

The overall idea of this project with Vi-Technology is to work towards manufacturing processes where machines communicate automatically so as to optimize the process performance as a whole. Starting from the assumption that transmitted information is essentially of statistical nature, the role of MISTIS in this context was to identify what statistical methods might be useful for the printed circuits boards assembly industry. A first step was to extract and analyze data from two inspection machines in an industrial process making electronic cards. After a first extraction in the SQL database, the goal was to enlighten the statistical links between these machines. Preliminary experiments and results on the Solder Paste Inspection (SPI) step, at the beginning of the line, helped identifying potentially relevant variables and measurements (eg related to stencil offsets) to identify future defects and discriminate between them. More generally, we had access to two databases at both ends (SPI and Component Inspection) of the assembly process. The goal was to improve our understanding of interactions in the assembly process, find out correlations between defects and physical measures, generate proactive alarms so as to detect departures from normality.

### *7.2.10. Tracking and analysis of large population of dynamic single molecules*

**Participant:** Florence Forbes.

**Joint work with**: Virginie Stoppin-Mellet from Grenoble Institute of Neuroscience, Vincent Brault from Laboratoire Jean Kuntzmann, Emilie Lebarbier from Nanterre University and Guy Bendao from AgroParisTech.

In the last decade, the number of studies using single molecule approaches has increased significantly. Thanks to technological progress and in particular with the development of TIRFM (Total Internal Reflection Fluorescence Microscopy), biologists can now observe single molecules at work. However, real time single molecule approaches remain mastered by a limited number of labs, and challenging obstacles have to be overcome before it becomes more broadly accessible. One important issue is the efficient detection and tracking of individual molecules in noisy images (low signal-to-noise ratio, SNR). Considering for example a TIRFM movie where single molecules stochastically appear and disappear at random positions, the low SNR implies that each individual molecule has to be detected at sub-pixel resolution over its local background and that this operation has to be repeated on each frame of the movie, thus requiring considerable amount of calculations. Procedures to detect single molecules are available, but they are mostly applicable to immobile molecules, are not statistically robust, and they often require an image processing that alters the quantitative signal information. In particular the intensity of a signal might be modified so that it becomes difficult to know the number of molecules associated with a specific signal. Crucial information such as the stoichiometry of the molecular complexes are then lost. Another challenging issue concerns data processing. Molecule tracking generate traces of time-dependent intensity fluctuations for each molecule. But single traces contain limited amount of information, and thus a very large number of traces must be analysed to extract general rules. In this context, the first aim of the present project was to provide a general procedure to track in real time transient interactions of a large number of biological molecules observed with TIRF microscopy and to generate traces of time-dependent intensity fluctuations. The second aim was to define a robust statistical approach to detect

discrete events in a noisy time-dependent signal and extract parameters that describe the kinetics of these events. For this task we gathered expertise from biology (Grenoble Institute of Neuroscience) and statistics (Inria Mistis, LJK and AgroParisTech) in the context of a multidisciplinary project funded by the Grenoble data institute for 2 years.

## 7.3. Graphical and Markov models

### *7.3.1. Structure learning via Hadamard product of correlation and partial correlation matrices*

**Participants:** Sophie Achard, Karina Ashurbekova, Florence Forbes.

Structure learning is an active topic nowadays in different application areas, i.e. genetics, neuroscience. Classical conditional independences or marginal independences may not be sufficient to express complex relationships. This work [39] is introducing a new structure learning procedure where an edge in the graph corresponds to a non zero value of both correlation and partial correlation. Based on this new paradigm, we define an estimator and derive its theoretical properties. The asymptotic convergence of the proposed graph estimator and its rate are derived. Illustrations on a synthetic example and application to brain connectivity are displayed.

### *7.3.2. Optimal shrinkage for robust covariance matrix estimators in a small sample size setting*

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

When estimating covariance matrices, traditional sample covariance-based estimators are straightforward but suffer from two main issues: 1) a lack of robustness, which occurs as soon as the samples do not come from a Gaussian distribution or are contaminated with outliers and 2) a lack of data when the number of parameters to estimate is too large compared to the number of available observations, which occurs as soon as the covariance matrix dimension is greater than the sample size. The first issue can be handled by assuming samples are drawn from a heavy-tailed distribution, at the cost of more complex derivations, while the second issue can be addressed by shrinkage with the difficulty of choosing the appropriate level of regularization. In this work [66] we offer both a tractable and optimal framework based on shrinked likelihood-based M-estimators. First, a closed-form expression is provided for a regularized covariance matrix estimator with an optimal shrinkage coefficient for any sample distribution in the elliptical family. Then, a complete inference procedure is proposed which can also handle both unknown mean and tail parameter, in contrast to most existing methods that focus on the covariance matrix parameter requiring pre-set values for the others. An illustration on synthetic and real data is provided in the case of the t-distribution with unknown mean and degrees-of-freedom parameters.

### *7.3.3. Robust penalized inference for Gaussian Scale Mixtures*

**Participants:** Sophie Achard, Karina Ashurbekova, Florence Forbes.

The literature on sparse precision matrix estimation is rapidly growing. Many strong methods are valid only for Gaussian variables. One of the most commonly used approaches in this case is glasso which aims to minimize the negative L1-penalized log-likelihood function. In practice, data may deviate from normality in various ways, outliers and heavy tails frequently occur that can severely degrade the Gaussian models performance. A natural solution is to turn to heavier tailed distributions that remain tractable. For this purpose, we propose [51] a penalized version of the EM algorithm for Gaussian Scale Mixtures.

### *7.3.4. Non parametric Bayesian priors for graph structured data*

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

We consider the issue of determining the structure of clustered data, both in terms of finding the appropriate number of clusters and of modelling the right dependence structure between the observations. Bayesian nonparametric (BNP) models, which do not impose an upper limit on the number of clusters, are appropriate to avoid the required guess on the number of clusters but have been mainly developed for independent data. In contrast, Markov random fields (MRF) have been extensively used to model dependencies in a tractable manner but usually reduce to finite cluster numbers when clustering tasks are addressed. Our main contribution is to propose a general scheme to design tractable BNP-MRF priors that combine both features: no commitment to an arbitrary number of clusters and a dependence modelling. A key ingredient in this construction is the availability of a stick-breaking representation which has the threefold advantage to allowing us to extend standard discrete MRFs to infinite state space, to design a tractable estimation algorithm using variational approximation and to derive theoretical properties on the predictive distribution and the number of clusters of the proposed model. This approach is illustrated on a challenging natural image segmentation task for which it shows good performance with respect to the literature. This work [77] will be presented as a poster at BayesComp2020 in Gainesville, Florida, USA, [78].

### *7.3.5. Bayesian nonparametric models for hidden Markov random fields on count variables and application to disease mapping*

**Participants:** Julyan Arbel, Fatoumata Dama, Jean-Baptiste Durand, Florence Forbes.

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 [77] has focused on Bayesian nonparametric priors for HMRFs and continuous, Gaussian observations. In this work, we consider extensions to discrete observed data typically issued from counts. We define and implement Bayesian nonparametric models for HMRFs with Poisson distributed observations. As an illustration, we propose a new disease mapping model for epidemiology. The inference is done by Variational Bayesian Expectation Maximization (VBEM). Results on synthetic data sets suggest that our model is able to recover the true number of risk levels (clusters) and to provide a good estimation of the true risk level partition. Application on real data then also shows satisfying results.

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.

### *7.3.6. Hidden Markov models for the analysis of eye movements*

**Participants:** Jean-Baptiste Durand, Brice Olivier, Sophie Achard.

*This research theme is supported by a LabEx PERSYVAL-Lab project-team grant.*
**Joint work with**: Anne Guérin-Dugué (GIPSA-lab) and Benoit Lemaire (Laboratoire de Psychologie et Neurocognition)

In the last years, GIPSA-lab has developed computational models of information search in web-like materials, using data from both eye-tracking and electroencephalograms (EEGs). These data were obtained from experiments, in which subjects had to decide whether a text was related or not to a target topic presented to them beforehand. In such tasks, reading process and decision making are closely related. Statistical analysis of such data aims at deciphering underlying dependency structures in these processes. Hidden Markov models (HMMs) have been used on eye-movement series to infer phases in the reading process that can be interpreted as strategies or steps in the cognitive processes leading to decision. In HMMs, each phase is associated with a state of the Markov chain. The states are observed indirectly though eye-movements. Our approach was

inspired by Simola *et al.* (2008) [86], but we used hidden semi-Markov models for better characterization of phase length distributions (Olivier *et al.*, 2017) [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. A comparison between the effects of three types of texts was performed, considering texts either closely related, moderately related or unrelated to the target topic.

Then, using the restored state values, statistical characteristics of EEGs were compared according to strategies, brain wave frequencies and EEG channels (i.e., location on scalp). Differences in variance and correlations related to strategy changes were highlighted. Dependency graphs interpreted as maps of functional brain connectivity were estimated for each strategy and frequency and their changes were interpreted.

These results were published in Brice Olivier's PhD manuscript [12]. Although the approach was sufficient to highlight significant discrimination of strategies, it suffered from somewhat overlapping eye-movement characteristics over strategies. As a result, high uncertainty in the phase changes arose, which could induce underestimation of EEG and eye movement abilities to discriminate strategies.

This is why we developed integrated models coupling EEG and eye movements within one single HMM for better identification of strategies. Here, the coupling incorporated some delay between transitions in both EEG and eye-movement state sequences, since EEG patterns associated to cognitive processes occur lately with respect to eye-movement state switches. Moreover, EEGs and scanpaths were recorded with different time resolutions, so that some resampling scheme had to be added into the model, for the sake of synchronizing both processes. An associated EM algorithm for maximum likelihood parameter estimation was derived.

Our goal for this coming year is to implement and validate our coupled model for jointly analyzing eye-movements and EEGs in order to improve the discrimination of reading strategies.

### *7.3.7. Comparison of initialization strategies in the EM algorithm for hidden Semi-Markov processes*

**Participants:** Jean-Baptiste Durand, Brice Olivier.

*This research theme is supported by a LabEx PERSYVAL-Lab project-team grant.*
**Joint work with**: Anne Guérin-Dugué (GIPSA-lab)

In Subsection 7.3.6, hidden semi-Markov models (HSMMs) were used to infer reading strategies from eye-movement and EEG time series. Model parameters were estimated by the EM algorithm. Its principle is to build a sequence of parameters with increasing likelihood values, starting from a starting point. The impact of this starting point has not been investigated in the case of HSMMs; this is why we aimed at developing and assessing an initialization method based on the available sequence lengths [48]. This consists in randomly choosing a number of transitions and then, uniformly-distributed transition times given the number of transitions. These transition times break the sequences into segements and assign uniformly-distributed states to each segment with the constraint that two consecutive states should be different.

The method was compared to other initialization strategies and was shown to be efficient on several data sets with multiple categorical sequences.

### *7.3.8. Lossy compression of tree structures*

**Participant:** Jean-Baptiste Durand.

**Joint work with**: Christophe Godin and Romain Azaïs (Inria Mosaic)

The class of self-nested trees presents remarkable compression properties because of the systematic repetition of subtrees in their structure. The aim of our work is to achieve compression of any unordered tree by finding the nearest self-nested tree. Solving this optmization problem without more assumptions is conjectured to be an NP-complete or NP-hard problem. In [40], we firstly provided a better combinatorial characterization of this specific family of trees. In particular, we showed from both theoretical and practical viewpoints that complex queries can be quickly answered in self-nested trees compared to general trees. We also presented an approximation algorithm of a tree by a self-nested one that can be used in fast prediction of edit distance between two trees.

Our goal for this coming year is to apply this approach to quantify the degree of self-nestedness of several plant species and extend first results obtained on rice panicles stating that near self-nestedness is a fairly general pattern in plants.

#### *7.3.9. Bayesian neural networks*

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

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

We investigate in [45] deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities, shedding light on novel sparsity-inducing mechanisms at the level of the units of the network, both pre- and post-nonlinearities. The main thrust of the paper is to establish that the units prior distribution becomes increasingly heavy-tailed with depth. We show that first layer units are Gaussian, second layer units are sub-Exponential, and we introduce sub-Weibull distributions to characterize the deeper layers units. Bayesian neural networks with Gaussian priors are well known to induce the weight decay penalty on the weights. In contrast, our result indicates a more elaborate regularisation scheme at the level of the units. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.

# 8. Bilateral Contracts and Grants with Industry

## 8.1. Bilateral Contracts with Industry

**Contract with EDF (2019).** Stéphane Girard is the advisor of the internship of Valentin Chevalier founded by EDF. The goal is to investigate sensitivity analysis and extrapolation limits in extreme-value theory with application to extreme weather events. The financial support for MISTIS is of 50 keuros.

**Contract with VALEO (2018-2019).** Stéphane Girard and Pascal Dkengne Sielenou are involved in a study with Valeo to assess the relevance of extreme-value theory in the calibration of sensors for autonomous cars. The financial support for MISTIS is of 100 keuros.

**Contract with Andritz.** F. Forbes and C. Braillon (SED) are involved in a study with Andritz to elaborate metrics based on image analysis to assess the quality of nonwaven tissues. The financial support for MISTIS is of 15 keuros.

# 9. Partnerships and Cooperations

## 9.1. National Initiatives

### *9.1.1. ANR*

MISTIS 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 MISTIS is about 15.000 euros.

### 9.1.2. Grenoble Idex projects

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

- The main objective of the RISK@UGA project is to provide some innovative tools both for the management of risk and crises in areas that are made vulnerable because of strong interdependencies between human, natural or technological hazards, in synergy with the conclusions of Sendai conference. The project federates a hundred researchers from Human and Social Sciences, Information & System Sciences, Geosciences and Engineering Sciences, already strongly involved in the problems of risk assessment and management, in particular natural risks. The PhD thesis of Meryem Bousebata is one of the eleven PhDs funded by this project.
- The NeuroCoG project aims at understanding the biological, neurophysiological and functional bases of behavioral and cognitive processes in normal and pathological conditions, from cells to networks and from individual to social cognition. No decisive progress can be achieved in this area without an aspiring interdisciplinary approach. The interdisciplinary ambition of NeuroCoG is particularly strong, bringing together the best scientists, engineers and clinicians at the crossroads of experimental and life sciences, human and social sciences and information and communication sciences, to answer major questions on the workings of the brain and of cognition. One of the work package entitled InnobioPark is dedicated to Parkinson's Disease. The PhD thesis of Veronica Munoz Ramirez is one of the three PhDs in this work package.
- The Grenoble Alpes Data Institute aims at undertaking groundbreaking interdisciplinary research focusing on how data change science and society. It combines three fields of data-related research in a unique way: data science applied to spatial and environmental sciences, biology, and health sciences; data-driven research as a major tool in Social Sciences and Humanities; and studies about data governance, security and the protection of data and privacy. In this context, a 2-year multi-disciplinary projects has been granted in November 2018 to Mistis in collaboration with the Grenoble Institute of Neuroscience. The objective of this project is to develop a statistical learning technique that is able to solve a problem of tracking and analyzing a large population of single molecules. The main difficulties are: 1) the large number of observations to analyse, 2) the noisy nature of the signals, 3) the definition of a quality index to allow the elimination of poor-quality data and false positive signals. We also aim at providing a powerful, well-documented and open-source software, that will be user-friendly for non-specialists.

Also 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.

### 9.1.3. Competitivity Clusters

**The MINALOGIC VISION 4.0 project:** MISTIS is involved in a three-year (2016-19) project. The project is led by VI-Technology, a world leader in Automated Optical Inspection (AOI) of a broad range of electronic components. The other partners are the G-Scop Lab in Grenoble and ACTIA company based in Toulouse. Vision 4.0 (in short Vi4.2) is one of the 8 projects labeled by Minalogic, the digital technology competitiveness cluster in Auvergne-Rhône-Alpes, that has been selected for the Industry 4.0 topic in 2016, as part of the 22nd call for projects of the FUI-Régions, for a total budget of the project of 3,4 Meuros.

Today, in the printed circuits boards (PCB) assembly industry, the assembly of electronic cards is a succession of ultra automated steps. Manufacturers, in constant quest for productivity, face sensitive and complex adjustments to reach ever higher levels of quality. Project VI4.2 proposes to build an innovative software solution to facilitate these adjustments, from images and measures obtained in automatic optical inspection (AOI). The idea is - from a centralized station for all the assembly line devices - to analyze and model the defects finely, to adjust each automatic machine, and to configure the interconnection logic between them to improve the quality. Transmitted information is essentially of statistical nature and the role of sc mistis is to identify which statistical methods might be useful to exploit at best the large amount of data registered by

AOI machines. Preliminary experiments and results on the Solder Paste Inspection (SPI) step, at the beginning of the assembly line, helped determining candidate variables and measurements to identify future defects and to discriminate between them. More generally, the idea is to analyze two databases at both ends (SPI and Component Inspection) of the assembly process so as to improve our understanding of interactions in the assembly process, find out correlations between defects and physical measures and generate accordingly proactive alarms so as to detect as early as possible departures from normality.

### *9.1.4. Networks*

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

## 9.2. European Initiatives

### *9.2.1. FP7 & H2020 Projects*

**VHIA ERC project (2015-19).**

MISTIS is involved in R. Horaud's ERC advanced Grant entitled Vision and Hearing In Action. VHIA studies the fundamentals of audio-visual perception for human-robot interaction.

## 9.3. International Initiatives

### *9.3.1. Inria International Labs*

**International Laboratory for Research in Computer Science and Applied Mathematics**
Associate Team involved in the International Lab:

#### *9.3.1.1. SIMERG2E*

Title: Statistical Inference for the Management of Extreme Risks, Genetics and Global Epidemiology

International Partner (Institution - Laboratory - Researcher):

UGB (Senegal) Abdou Kâ Diongue

Start year: 2018

See also: http://mistis.inrialpes.fr/simerge

SIMERG2E is built on the same two research themes as SIMERGE, with some adaptations to new applications: 1) Spatial extremes, application to management of extreme risks. We address the definition of new risk measures, the study of their properties in case of extreme events and their estimation from data and covariate information. Our goal is to obtain estimators accounting for possible variability, both in terms of space and time, which is of prime importance in many hydrological, agricultural and energy contexts. 2) Classification, application to genetics and global epidemiology. We address the challenge to build statistical models in order to test association between diseases and human host genetics in a context of genome-wide screening. Adequate models should allow to handle complexity in genomic data (correlation between genetic markers, high dimensionality) and additional statistical issues present in data collected from a family-based longitudinal survey (non-independence between individuals due to familial relationship and non-independence within individuals due to repeated measurements on a same person over time).

### *9.3.2. Inria Associate Teams Not Involved in an Inria International Labs*

#### *9.3.2.1. LANDER*

Title: Latent Analysis, Adversarial Networks, and DimEnsionality Reduction

International Partner (Institution - Laboratory - Researcher):

La Trobe university, Melbourne (Australia) - Department of Mathematics - Hien Nguyen

Start year: 2019

See also: https://team.inria.fr/mistis/projects/lander/

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

### *9.3.3. Inria International Partners*

#### *9.3.3.1. Informal International Partners*

The context of our research is also the collaboration between MISTIS and a number of international partners such as the statistics department of University of Michigan, in Ann Arbor, USA, the statistics department of McGill University in Montreal, Canada, Université Gaston Berger in Senegal and Universities of Melbourne and Brisbane in Australia.

The main other active international collaborations in 2019 are with:

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

## 9.4. International Research Visitors

### *9.4.1. Visits of International Scientists*

- Bernardo Nipoti, assistant professor at Milano Bicocca University, Italy, visited for a month in 2019 (three visits in February, April and September).
- Natalie Karavarsamis, assistant professor at La Trobe University in Melbourne, Australia, visited for a week in November 2019.
- Hien Nguyen, researcher at La Trobe University in Melbourne, Australia, visited for a month in November 2019.
- Darren Wraith, assistant professor at QUT, Brisbane, Australia, visited for 2 weeks in December 2019 and January 2020.
- Aboubacrène Ag Ahmad, PhD student at Univ. Gaston Berger, Senegal visited from September 2019 until November 2019.

*9.4.1.1. Internships*

Sharan Yalburgi did an internship of three months with Julyan Arbel on *Bayesian deep learning for model selection and approximate inference*.

*9.4.1.2. Research Stays Abroad*

Mariia Vladimirova visited David Dunson at Duke University for three months (Nov 2019 - Jan 2020).

# 10. Dissemination

## 10.1. Promoting Scientific Activities

### *10.1.1. Scientific Events Organisation*

*10.1.1.1. General Chair, Scientific Chair*

- Stéphane Girard was chairman at the 2nd workshop on Multivariate Data and Software (Limassol, Cyprus) and at the International Workshop on Stress Test and Risk Management (Paris).

*10.1.1.2. Member of the Organizing Committees*

- Florence Forbes is a member of the scientific committees of the Bayes Comp 2020 conference in Gainesville, Florida, USA (January 2020) and of the Research school on Networks and molecular biology at CIRM in Marseille (March 2020).
- Sophie Achard was a member of the scientific committee of the Wavelet & Sparsity XVIII 2019 in San Diego and the organizer of a special session within this conference.
- Stéphane Girard and Julyan Arbel were members of the organizing committee of the 10th Statlearn international workshop "Challenging problems in Statistical Learning", Grenoble, http://statlearn.sfds.asso.fr. Stéphane Girard also co-organized (with D. Fraix-Burnet, IPAG) the 4th international school Stat4Astro "Variability and Time Series Analysis", Autrans, http://stat4astro2019.sciencesconf.org.
- Julyan Arbel was a member of the scientific committee of *Statistical Methods for Post Genomic Data analysis (SMPGD)*, link. Julyan Arbel organized the session entitled 'Bayesian Machine Learning' at the 12th International Conference of Computational and Methodological Statistics (CMStat), University of London, UK (14-16 December 2019).

**Seminars organization**

- MISTIS participates in the weekly statistical seminar of Grenoble. Several lecturers have been invited in this context.
- Julyan Arbel is organizing monthly reading group Bayes in Grenoble on Bayesian statistics.

### 10.1.2. Scientific Events Selection

*10.1.2.1. Reviewer*

- In 2019, Florence Forbes has been a reviewer for CAP 2019 in Toulouse and for ICDHT 2019 in Tunis.
- In 2019, Julyan Arbel has been a reviewer for the *Bayesian Young Statisticians Meeting proceedings* (BAYSM).
- In 2019, Florence Forbes and Julyan Arbel have been reviewers for the *Research School on Statistics and Data Science* (RSSDS2019).

### 10.1.3. Journal

*10.1.3.1. Member of the Editorial Boards*

- Stéphane Girard is Associate Editor of the *Statistics and Computing* journal since 2012, Associate Editor of the *Journal of Multivariate Analysis* since 2016 and Associate Editor of *REVSTAT - Statistical Journal* since 2019. He is also member of the Advisory Board of the *Dependence Modelling* journal since December 2014.
- Florence Forbes is Associate Editor of the journal *Frontiers in ICT: Computer Image Analysis* since its creation in Sept. 2014. She is also Associate Editor of the *Computational Statistics and Data Analysis* journal since May 2018.
- Julyan Arbel is Associate Editor of *Bayesian Analysis* (BA) and of Statistics & Probability Letters (SPL) since 2019.
- Julyan Arbel and Florence Forbes are Associate Editors for the *Australian & New Zealand Journal of Statistics* (ANZJS), since 2018.
- Sophie Achard is Associate Editor of *Neural Processing Letters* and *Network Neuroscience* since 2016.

*10.1.3.2. Reviewer - Reviewing Activities*

- In 2019, Florence Forbes has been a reviewer for *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), *Statistics and Computing* (STCO), and *Neural Processing Letters*.
- In 2019, Stéphane Girard has been a reviewer for *Journal of the American Statistical Association* (JASA), *Journal of Statistical Planning and Inference* (JSPI), *Communications in Statistics - Theory and Methods, Spatial Statistics.*
- In 2019, Jean-Baptiste Durand has been a reviewer for *Behavior Research Methods* (BRM) and *Statistics and Computing* (STCO).
- In 2019, Julyan Arbel has been a reviewer for: *Annals of Applied Statistics* (AOAS), *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* (AIHP), *Bernoulli*, *Biometrika*, *Entropy*, *Journal of the American Statistical Association* (JASA), *Journal of Computational and Graphical Statistics* (JCGS), *Journal of Nonparametric Statistics* (JNS), *Sankhyā*, *Stats*, *Statistica Sinica*, *Statistics and Probability Letters* (SPL), *Stochastic Processes and their Applications* (SPA), *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI).

### 10.1.4. Invited Talks

Florence Forbes has been invited to give talks at the following seminars and conferences:

- Glasgow Statistics Department, March 2019
- Conference on Applied Inverse Problems, AIP 2019, July 2019 [34]
- 51ème Journées de la Statistique, Nancy, France, June 2019 [35]
- Workshop on Model-based clustering, Vienna, Austria, July 15-19, [32]
- Research School on Statistics and Data Science, Melbourne, Australia, July 24-26, [33]

Julyan Arbel has been invited to give talks at the following seminars and conferences:

- Applied Inverse Problems conference, Grenoble, France, July 8-12, 2019. Invited talk: Understanding Priors in Bayesian Neural Networks at the Unit Level.
- 11th Workshop on Bayesian Inference in Stochastic Processes (BISP), Madrid, Spain, June 12-14, 2019. Invited talk: Understanding Priors in Bayesian Neural Networks at the Unit Level.
- Workshop on Multivariate Data Analysis, Limassol, Cyprus, April 14-16, 2019. Invited talk: Some distributional properties of Bayesian neural networks.
- Seminar, Laboratoire d'Informatique de Grenoble, Grenoble, France, September 19, 2019. Invited talk: On some theory for Bayesian neural networks.
- Grenoble R User Group, Grenoble, France, April 11, 2019. Invited talk: R Markdown.

Sophie Achard has been invited to give talks at the following seminars and conferences:

- Workshop ATLAS, GDR MADICS, November 2019, Grenoble http://ama.liglab.fr/ATLAS/Wksp-22112019.html. Invited talks: Assessing reliability of resting-state fMRI graph analysis: challenges in measuring brain connectivity networks alterations for clinical applications.
- NeuroSTIC, GDR BioComp et ISIS, October 2019, Nice http://www.gdr-isis.fr/neurostic/?p=452. Invited talks: Brain connectivity for patients with consciousness disorders: statistical and clinical challenges

Stéphane Girard has been invited to give talks at the following seminars and conferences:

- 2nd workshop on Multivariate Data and Software (Limassol, Cyprus) [37],
- Workshop "Appréhender la grande dimension" (Paris) [36],
- Seminar, Nottingham University, UK "Estimation of extreme risk measures based on $L_p$- quantiles".

Antoine Usseglio-Carleve was invited to give a talk [38] at the 12th International Conference of Computational and Methodological Statistics, London, UK.

Marta Crispino was invited to give a talk [31] at the 12th International Conference of Computational and Methodological Statistics, London, UK.

### *10.1.5. Scientific Expertise*

Florence Forbes is Scientific Advisor since March 2015 for the Pixyl company.

### *10.1.6. Research Administration*

- Stéphane Girard is a member of the "Comité des Emplois Scientifiques" at Inria Grenoble Rhône-Alpes since 2015.
- Since 2015, Stéphane Girard is a member of the INRA committee (CSS MBIA) in charge of evaluating INRA researchers once a year in the MBIA dept of INRA.
- Florence Forbes is a member of the "Comité Développement Technologique" for software development projects at Inria Grenoble Rhône-Alpes since 2015.
- Florence Forbes is a member of the "Comite d'organisation stratégique" of Inria Grenoble Rhône-Alpes since 2017.
- Florence Forbes is a member of the Executive Committee of the Grenoble data institute.
- Florence Forbes has been a member of the Selection committee for assistant professors at Ensimag Grenoble and at Ecole Centrale Lille in 2019.
- Florence Forbes is a member of the advisory committee of the Helmholtz AI Cooperation Unit https://helmholtz.ai/, since 2019.
- Sophie Achard is co-director of pôle MSTIC whithin Université Grenoble Alpes, since 2017.
- Julyan Arbel is a scientific committee member of the Data Science axis of Persyval Labex (Machine learning: fundamentals and applications, and Data linking, sharing and privacy), link, since 2019.

## 10.2. Teaching - Supervision - Juries

### *10.2.1. Teaching*

- Master : Stéphane Girard, *Statistique Inférentielle Avancée*, 18 ETD, M1 level, Ensimag. Grenoble-INP, France.
- PhD course: Julyan Arbel, *Bayesian nonparametrics*, Jyväskylä Summer School, Finland, August 2019, 25 ETD.
- Master and PhD course: Julyan Arbel, *Bayesian statistics*, Ensimag, Université Grenoble Alpes (UGA), 25 ETD.
- Master and PhD course: Julyan Arbel, *Bayesian nonparametric statistics*, Master Mathématiques Apprentissage et Sciences Humaines (M*A*S*H), Université Paris-Dauphine, 25 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: Karina Ashurbekova "*Robust Structure Learning*", December 2019, Sophie Achard and Florence Forbes, Université Grenoble Alpes.
- PhD defended: Brice Olivier "*Joint analysis of eye-movements and EEGs using coupled hidden Markov models*", June 2019, Jean-Baptiste Durand and Anne Guérin-Dugué, Université Grenoble Alpes.
- PhD defended: Chun-Chen Tu,"*Gaussian mixture sub-clustering/reduction refinement of Non-linear high-to-low dimensional mapping*", "Date", Florence Forbes and Naisyin Wang, University of Michigan, Ann Arbor.
- HDR: Julyan Arbel, Université Grenoble Alpes, "*Bayesian Statistical Learning and Applications*" [11], October 2019.
- PhD in progress: Veronica Munoz,"*Extraction de signatures dans les données IRM de patients parkinsoniens de novo*", Florence Forbes and Michel Dojat, Université Grenoble Alpes, started on October 2017.
- PhD in progress: Fabien Boux,"*Développement de méthodes statistiques pour l'imagerie IRM fingerprinting*", Florence Forbes and Emmanuel Barbier, Université Grenoble Alpes, started on October 2017.
- PhD in progress: Benoit Kugler, "*Massive hyperspectral images analysis by inverse regression of physical models*", Florence Forbes and Sylvain Douté, Université Grenoble Alpes, started on October 2018.
- PhD in progress: Mariia Vladimirova, "*Prior specification for Bayesian deep learning models and regularization implications*", started on October 2018, Julyan Arbel and Jakob Verbeek.
- PhD in progress: Aboubacrène Ag Ahmad "*A new location-scale model for heavy-tailed distributions*", started on September 2016, Stéphane Girard and Alio Diop (Université Gaston Berger, Sénégal).
- PhD in progress: Meryem Bousebata "*Bayesian estimation of extreme risk measures: Implication for the insurance of natural disasters*", started on October 2018, Stéphane Girard and Geffroy Enjolras (Université Grenoble Alpes).

- PhD in progress: Alexandre Constantin "*Analyse de séries temporelles massives d'images satellitaires : Applications à la cartographie des écosystèmes*", started on November 2018, Stéphane Girard and Mathieu Fauvel (Université Grenoble Alpes).
- PhD in progress: Daria Bystrova, "*Joint Species Distribution Modeling: Dimension reduction using Bayesian nonparametric priors*", started on October 2019, Julyan Arbel and Wilfried Thuiller.
- PhD in progress: Giovanni Poggiatto , "*Scalable Approaches for Joint Species Distribution Modeling*", started on November 2019, Julyan Arbel and Wilfried Thuiller.

### *10.2.3. Juries*

- Julyan Arbel has been reviewer for the PhD thesis of Romain Mismer, LPSM, Sorbonne Université, Paris.
- Stéphane Girard has been reviewer for the PhD thesis of Maxime Baelde, Université de Lille.
- Stéphane Girard has been the president of the HDR committee of Julie Carreau, Université de Montpellier, and an examinator for the HDR of Julyan Arbel.
- Stéphane Girard has been a member of the PhD committee of Abdul-Fattah Abu-Awwad, Université de Lyon.
- Florence Forbes has been reviewer for the PhD thesis of Lê-Huu D. Khuê, Université Paris Saclay, CentraleSupelec, Cedric Meurée, Université de Rennes, Bao Tuyen Huynh Université de Caen and for the HDR thesis of Christine Keribin, Université Paris Orsay.
- Florence Forbes has been a member of the PhD committee of Charlotte Maugard, Université Grenoble Alpes and Esteban Bautista, ENS Lyon.
- Sophie Achard has been reviewer for the HDR of Julien Modolo, Université Rennes 1.

## 10.3. Popularization

### *10.3.1. Interventions*

- Sophie Achard has been invited to Festival des Nouvelles Explorations https://nouvellesexplorations.com/.
- Julyan Arbel gave a presentation for ISN conference, March 2019.

# 11. Bibliography

## Major publications by the team in recent years

[1] C. AMBLARD, S. GIRARD. *Estimation procedures for a semiparametric family of bivariate copulas*, in "Journal of Computational and Graphical Statistics", 2005, vol. 14, n$^{o}$ 2, pp. 1–15

[2] J. BLANCHET, F. FORBES. *Triplet Markov fields for the supervised classification of complex structure data*, in "IEEE trans. on Pattern Analyis and Machine Intelligence", 2008, vol. 30(6), pp. 1055–1067

[3] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional data clustering*, in "Computational Statistics and Data Analysis", 2007, vol. 52, pp. 502–519

[4] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional discriminant analysis*, in "Communication in Statistics - Theory and Methods", 2007, vol. 36, n$^{o}$ 14

[5] L. CHAARI, T. VINCENT, F. FORBES, M. DOJAT, P. CIUCIU. *Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach*, in "IEEE Transactions on Medical Imaging", May 2013, vol. 32, n° 5, pp. 821-837 [*DOI :* 10.1109/TMI.2012.2225636], http://hal.inria.fr/inserm-00753873

[6] A. DAOUIA, S. GIRARD, G. STUPFLER. *Estimation of Tail Risk based on Extreme Expectiles*, in "Journal of the Royal Statistical Society series B", 2018, vol. 80, pp. 263–292

[7] A. DELEFORGE, F. FORBES, R. HORAUD. *High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables*, in "Statistics and Computing", February 2014 [*DOI :* 10.1007/s11222-014-9461-5], https://hal.inria.fr/hal-00863468

[8] F. FORBES, G. FORT. *Combining Monte Carlo and Mean field like methods for inference in hidden Markov Random Fields*, in "IEEE trans. Image Processing", 2007, vol. 16, n° 3, pp. 824-837

[9] F. FORBES, D. WRAITH. *A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering*, in "Statistics and Computing", November 2014, vol. 24, n° 6, pp. 971-984 [*DOI :* 10.1007/s11222-013-9414-4], 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", 2004, vol. 33, n° 2, pp. 205–234

# Publications of the year

## Doctoral Dissertations and Habilitation Theses

[11] J. ARBEL. *Bayesian Statistical Learning and Applications*, Université grenoble Alpes, CNRS, Institut des Géosciences et de l'Environnement, October 2019, Habilitation à diriger des recherches, https://tel.archives-ouvertes.fr/tel-02429156

[12] B. OLIVIER. *Joint analysis of eye movements and EEGs using coupled hidden Markov*, Université Grenoble Alpes, June 2019, https://tel.archives-ouvertes.fr/tel-02311373

## Articles in International Peer-Reviewed Journals

[13] A. A. AHMAD, E. H. DEME, A. DIOP, S. GIRARD. *Estimation of the tail-index in a conditional location-scale family of heavy-tailed distributions*, in "Dependence Modeling", 2019, vol. 7, pp. 394–417, https://hal.inria.fr/hal-02132976

[14] C. ALBERT, A. DUTFOY, L. GARDES, S. GIRARD. *An extreme quantile estimator for the log-generalized Weibull-tail model*, in "Econometrics and Statistics ", 2019, pp. 1-39, forthcoming [*DOI :* 10.1016/J.ECOSTA.2019.01.004], https://hal.inria.fr/hal-01783929

[15] C. ALBERT, A. DUTFOY, S. GIRARD. *Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles*, in "Extremes", 2019, forthcoming, https://hal.archives-ouvertes.fr/hal-01692544

[16] J. ARBEL, M. CRISPINO, S. GIRARD. *Dependence properties and Bayesian inference for asymmetric multivariate copulas*, in "Journal of Multivariate Analysis", November 2019, vol. 174, pp. 104530:1-20 [*DOI :* 10.1016/J.JMVA.2019.06.008], https://hal.archives-ouvertes.fr/hal-01963975

[17] J. ARBEL, P. DE BLASI, I. PRÜNSTER. *Stochastic approximations to the Pitman-Yor process*, in "Bayesian Analysis", June 2019, vol. 14, n$^{o}$ 3, pp. 753-771 [*DOI :* 10.1214/18-BA1127], https://hal.archives-ouvertes.fr/hal-01950654

[18] J. ARBEL, S. FAVARO. *Approximating predictive probabilities of Gibbs-type priors*, in "Sankhya A", September 2019, pp. 1-21, https://hal.archives-ouvertes.fr/hal-01693333

[19] J. ARBEL, O. MARCHAL, H. T. NGUYEN. *On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables*, in "ESAIM: Probability and Statistics", December 2019, https://hal.archives-ouvertes.fr/hal-01998252

[20] M. CRISPINO, E. ARJAS, V. VITELLI, N. BARRETT, A. FRIGESSI. *A Bayesian Mallows Approach to Non-Transitive Pair Comparison Data: How Human are Sounds?*, in "Annals of Applied Statistics", June 2019, vol. 13, n$^{o}$ 1, pp. 492-519 [*DOI :* 10.1214/18-AOAS1203], https://hal.archives-ouvertes.fr/hal-01972952

[21] A. DAOUIA, S. GIRARD, G. STUPFLER. *Extreme M-quantiles as risk measures: From L1 to Lp optimization*, in "Bernoulli", February 2019, vol. 25, n$^{o}$ 1, pp. 264-309 [*DOI :* 10.3150/17-BEJ987], https://hal.inria.fr/hal-01585215

[22] A. DAOUIA, S. GIRARD, G. STUPFLER. *Tail expectile process and risk assessment*, in "Bernoulli", 2019, pp. 1-27, forthcoming, https://hal.archives-ouvertes.fr/hal-01744505

[23] L. GARDES, S. GIRARD, G. STUPFLER. *Beyond tail median and conditional tail expectation: extreme risk estimation using tail Lp -optimisation*, in "Scandinavian Journal of Statistics", 2019, pp. 1-69, forthcoming [*DOI :* 10.1111/SJOS.12433], https://hal.inria.fr/hal-01726328

[24] M. JALBERT, F. ZHENG, A. WOJTUSCISZYN, F. FORBES, S. BONNET, K. SKAARE, P.-Y. BENHAMOU, S. LABLANCHE. *Glycemic variability indices can be used to diagnose islet transplantation success in type 1 diabetic patients*, in "Acta Diabetologica", October 2019, pp. 1-11 [*DOI :* 10.1007/s00592-019-01425-3], https://hal.archives-ouvertes.fr/hal-02328170

[25] C. LAWLESS, J. ARBEL. *A simple proof of Pitman-Yor's Chinese restaurant process from its stick-breaking representation*, in "Dependence Modeling", 2019, vol. 7, n$^{o}$ 1, pp. 45-52 [*DOI :* 10.1515/DEMO-2019-0003], https://hal.archives-ouvertes.fr/hal-01950653

[26] Q. LIU, M. CRISPINO, I. SCHEEL, V. VITELLI, A. FRIGESSI. *Model-based learning from preference data*, in "Annual Reviews of Statistics and its Application", March 2019, vol. 6, n$^{o}$ 1, pp. 329-354 [*DOI :* 10.1146/ANNUREV-STATISTICS-031017-100213], https://hal.archives-ouvertes.fr/hal-01972948

[27] H. D. NGUYEN, F. CHAMROUKHI, F. FORBES. *Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model*, in "Neurocomputing", November 2019, vol. 366, pp. 208-214 [*DOI :* 10.1016/J.NEUCOM.2019.08.014], https://hal.archives-ouvertes.fr/hal-02265793

[28] H. D. NGUYEN, F. FORBES, G. J. MCLACHLAN. *Mini-batch learning of exponential family finite mixture models*, in "Statistics and Computing", 2019, pp. 1-40, forthcoming, https://hal.archives-ouvertes.fr/hal-02415068

[29] C.-C. TU, F. FORBES, B. LEMASSON, N. WANG. *Prediction with high dimensional regression via hierarchically structured Gaussian mixtures and latent variables*, in "Journal of the Royal Statistical Society:

Series C Applied Statistics", 2019, pp. 1-23 [*DOI :* 10.1111/RSSC.12370], https://hal.archives-ouvertes.fr/hal-02263144

[30] F. ZHENG, M. JALBERT, F. FORBES, S. BONNET, A. WOJTUSCISZYN, S. LABLANCHE, P.-Y. BENHAMOU. *Characterization of Daily Glycemic Variability in Subjects with Type 1 Diabetes Using a Mixture of Metrics*, in "Diabetes Technology and Therapeutics", 2019, pp. 1-17, forthcoming [*DOI :* 10.1089/DIA.2019.0250], https://hal.archives-ouvertes.fr/hal-02415078

### Invited Conferences

[31] M. CRISPINO, S. GIRARD, J. ARBEL. *Dependence properties and Bayesian inference for asymmetric multivariate copulas*, in "CMStatistics 2019 - 12th International Conference of the ERCIM WG on Computational and Methodological Statistics", London, United Kingdom, December 2019, https://hal.archives-ouvertes.fr/hal-02413948

[32] F. FORBES, A. ARNAUD, B. LEMASSON, E. L. BARBIER. *Bayesian mixtures of multiple scale distributions*, in "2019 - 26th Summer Working Group on Model-Based Clustering", Vienna, Austria, July 2019, https://hal.archives-ouvertes.fr/hal-02423638

[33] F. FORBES, A. ARNAUD, B. LEMASSON, E. L. BARBIER. *Component elimination strategies to fit mixtures of multiple scale distributions*, in "RSSDS 2019 - Research School on Statistics and Data Science", Melbourne, Australia, Proceedings of the Research School on Statistics and Data Science 2019, July 2019, pp. 1-15, https://hal.archives-ouvertes.fr/hal-02415090

[34] F. FORBES, A. DELEFORGE, R. HORAUD, E. PERTHAME. *Robust non-linear regression approach for generalized inverse problems in a high dimensional setting*, in "AIP 2019 - Applied Inverse Problem conference", Grenoble, France, July 2019, https://hal.archives-ouvertes.fr/hal-02415115

[35] F. FORBES, D. WRAITH. *Robust mixture modelling using skewed multivariate distributions with variable amounts of tailweight*, in "JdS 2019 - 51èmes Journées de Statistique", Nancy, France, Proceedings des 51èmes Journées de Statistique 2019, June 2019, https://hal.archives-ouvertes.fr/hal-02423639

[36] S. GIRARD. *Un aperçu des méthodes statistiques pour la classification et la régression en grande dimension*, in "Workshop "Appréhender la grande dimension" 2019", Paris, France, June 2019, https://hal.inria.fr/hal-02149891

[37] S. GIRARD, G. STUPFLER. *Estimation of high-dimensional extreme conditional expectiles*, in "CRoNoS & MDA 2019 - Final CRoNoS meeting and 2nd workshop on Multivariate Data Analysis", Limassol, Cyprus, April 2019, https://hal.inria.fr/hal-02099370

[38] A. USSEGLIO-CARLEVE, S. GIRARD, G. STUPFLER. *Nonparametric extreme conditional expectile estimation*, in "CMStatistics 2019 - 12th International Conference of the ERCIM WG on Computational and Methodological Statistics", London, United Kingdom, December 2019, https://hal.archives-ouvertes.fr/hal-02413682

### International Conferences with Proceedings

[39] K. ASHURBEKOVA, S. ACHARD, F. FORBES. *Structure Learning via Hadamard Product of Correlation and Partial Correlation Matrices*, in "EUSIPCO 2019 - 27th European Signal Processing Conference", A Coruña, Spain, IEEE, September 2019, pp. 1-5 [*DOI :* 10.23919/EUSIPCO.2019.8902948], http://hal.univ-grenoble-alpes.fr/hal-02290847

[40] R. AZAÏS, J.-B. DURAND, C. GODIN. *Approximation of trees by self-nested trees*, in "ALENEX 2019 - Algorithm Engineering and Experiments", San Diego, United States, SIAM, 2019, pp. 39-53, https://arxiv.org/abs/1810.10860 [*DOI :* 10.1137/1.9781611975499.4], https://hal.archives-ouvertes.fr/hal-01294013

[41] P. BRUEL, S. QUINITO MASNADA, B. VIDEAU, A. LEGRAND, J.-M. VINCENT, A. GOLDMAN. *Auto-tuning under Tight Budget Constraints: A Transparent Design of Experiments Approach*, in "CCGrid 2019 - International Symposium in Cluster, Cloud, and Grid Computing", Larcana, Cyprus, May 2019, pp. 1-10 [*DOI :* 10.1109/CCGRID.2019.00026], https://hal.inria.fr/hal-02110868

[42] S. GIRARD, G. STUPFLER, A. USSEGLIO-CARLEVE. *Nonparametric extreme conditional expectile estimation*, in "EVA 2019 - 11th International Conference on Extreme Value Analysis", Zagreb, Croatia, July 2019, 1 p. , https://hal.inria.fr/hal-02186705

[43] V. MUÑOZ RAMÍREZ, F. FORBES, J. ARBEL, A. ARNAUD, M. DOJAT. *Quantitative MRI characterization of brain abnormalities in de novo Parkinsonian patients*, in "ISBI 2019 - IEEE International Symposium on Biomedical Imaging", Venice, Italy, Proceedings of IEEE International Symposium on Biomedical Imaging, April 2019, pp. 1-4 [*DOI :* 10.1109/ISBI.2019.8759544], https://hal.archives-ouvertes.fr/hal-01970682

[44] V. MUÑOZ RAMÍREZ, F. FORBES, P. COUPÉ, M. DOJAT. *No Structural Differences Are Revealed by VBM in 'de novo' Parkinsonian Patients*, in "MEDINFO 2019 - 17th World Congress On Medical And Health Informatics", Lyon, France, August 2019, pp. 268-272 [*DOI :* 10.3233/SHTI190225], https://hal.inria.fr/hal-02426273

[45] M. VLADIMIROVA, J. VERBEEK, P. MESEJO, J. ARBEL. *Understanding Priors in Bayesian Neural Networks at the Unit Level*, in "ICML 2019 - 36th International Conference on Machine Learning", Long Beach, United States, Proceedings of the 36th International Conference on Machine Learning, June 2019, vol. 97, pp. 6458-6467, https://arxiv.org/abs/1810.05193 - 10 pages, 5 figures, ICML'19 conference, https://hal.archives-ouvertes.fr/hal-02177151

### National Conferences with Proceedings

[46] C. ALBERT, A. DUTFOY, S. GIRARD. *Etude de l'erreur relative d'extrapolation associée à l'estimateur de Weissman pour les quantiles extrêmes*, in "JdS 2019 - 51èmes Journées de Statistique", Nancy, France, Société Française de Statistique, June 2019, pp. 1-6, https://hal.inria.fr/hal-02149905

[47] J.-B. DURAND. *Compétitions d'analyse des données à l'Université Grenoble Alpes : motivations, organisation et retours d'expérience*, in "CFIES 2019 - Colloque francophone international sur l'enseignement de la statistique", Strasbourg, France, September 2019, pp. 1-6, https://hal.inria.fr/hal-02298606

[48] B. OLIVIER, A. GUÉRIN-DUGUÉ, J.-B. DURAND. *Assessment of various initialization strategies for the Expectation-Maximization algorithm for Hidden Semi-Markov Models with multiple categorical sequences*, in "JdS 2019 - 51èmes Journées de Statistique", Vandœuvre-lès-Nancy, France, June 2019, pp. 1-7, https://hal.inria.fr/hal-02129122

[49] F. ZHENG, S. BONNET, F. FORBES, M. JALBERT, S. LABLANCHE, P.-Y. BENHAMOU. *Caractérisation de la variabilité glycémique par analyse statistique multivariée*, in "GRETSI 2019 - XXVIIème Colloque francophonede traitement du signal et des images", Lille, France, Proceedings du XXVIIème Colloque francophone de traitement du signal et des images, August 2019, pp. 1-4, https://hal.archives-ouvertes.fr/hal-02415082

[50] F. ZHENG, M. JALBERT, F. FORBES, S. BONNET, A. WOJTUSCISZYN, S. LABLANCHE, P.-Y. BENHAMOU. *Caractérisation de la variabilité glycémique journalière chez le patient avec diabète de type 1*, in "SFD 2019 - Congrès annuel de la Société Francophone du Diabète", Marseille, France, Proceedings du Congrès annuel de la Société Francophone du Diabète, March 2019, https://hal.archives-ouvertes.fr/hal-01971621

### Conferences without Proceedings

[51] K. ASHURBEKOVA, S. ACHARD, F. FORBES. *Robust penalized inference for Gaussian Scale Mixtures*, in "SPARS 2019 - Workshop on Signal Processing with Adaptive Sparse Structured Representations", Toulouse, France, July 2019, pp. 1-2, http://hal.univ-grenoble-alpes.fr/hal-02291576

[52] *Best Paper*
M. BOUSEBATA, G. ENJOLRAS, S. GIRARD. *Bayesian estimation of natural extreme risk measures. Application to agricultural insurance*, in "IDRiM 2019 - 10th conference of the international society for Integrated Disaster Risk Management", Nice, France, October 2019, https://hal.archives-ouvertes.fr/hal-02276292.

[53] F. BOUX, F. FORBES, J. ARBEL, E. L. BARBIER. *Dictionary learning via regression: vascular MRI application*, in "CNIV 2019 - 3e Congrès National d'Imagerie du Vivant", Paris, France, February 2019, pp. 1-12, https://hal.archives-ouvertes.fr/hal-02428647

[54] F. BOUX, F. FORBES, J. ARBEL, E. L. BARBIER. *Estimation de paramètres IRM en grande dimension via une régression inverse*, in "SFRMBM 2020 - 4e congrés de la Société Française de Résonance Magnétique en Biologie et Médecine", Strasbourg, France, March 2020, 1 p. , https://hal.archives-ouvertes.fr/hal-02428679

[55] A. CONSTANTIN, M. FAUVEL, S. GIRARD, S. IOVLEFF, Y. TANGUY. *Classification de Signaux Multidimensionnels Irrégulièrement Echantillonnés*, in "2019 - Journée Jeunes Chercheurs MACLEAN du GDR MADICS", Paris, France, December 2019, pp. 1-2, https://hal.archives-ouvertes.fr/hal-02394120

[56] P. S. DKENGNE, S. GIRARD, S. AHIAD. *Estimation of the extrapolation range associated with extreme-value models: Application to the assessment of sensors reliability*, in "EMS 2019 - 32nd European Meeting of Statisticians", Palerme, Italy, July 2019, https://hal.archives-ouvertes.fr/hal-02278051

[57] V. MUÑOZ RAMÍREZ, M. DOJAT, F. FORBES. *Mixture Models for the characterization of brain abnormalities in "de novo" Parkinsonian patients*, in "CNIV 2019 - 3e Congrès National d'Imagerie du Vivant", Paris, France, February 2019, pp. 1-16, https://hal.inria.fr/hal-02436886

[58] V. MUÑOZ RAMÍREZ, F. FORBES, A. ARNAUD, M. DOJAT. *Brain abnormalities detection in de Novo Parkinsonian patients*, in "OHBM 2019 - 25th Annual Meeting of the Organization for Human Brain Mapping", Rome, Italy, June 2019, pp. 1-11, https://hal.archives-ouvertes.fr/hal-02415101

[59] V. MUÑOZ RAMÍREZ, F. FORBES, P. COUPÉ, M. DOJAT. *No structural Brain differences in 'de novo' Parkinsonian patients*, in "OHBM 2019 - 25th Annual Meeting of the Organization for Human Brain Mapping", Rome, Italy, June 2019, pp. 1-5, https://hal.archives-ouvertes.fr/hal-02192447

[60] S. SALHI, F. BONNEFOY, S. GIRARD, M. BERNIER, N. BARBOT, R. SIRAGUSA, E. PERRET, F. GARET. *Enhanced THz tags authentication using multivariate statistical analysis*, in "IRMMW-THz 2019 - 44th International Conference on Infrared, Millimeter, and Terahertz Waves", Paris, France, September 2019, pp. 1-2, https://hal.archives-ouvertes.fr/hal-02282841

### Scientific Books (or Scientific Book chapters)

[61] K. K. MENGERSEN, E. DUNCAN, J. ARBEL, C. ALSTON-KNOX, N. WHITE. *Applications in Industry*, in "Handbook of mixture analysis", S. FRUHWIRTH-SCHNATTER, G. CELEUX, C. P. ROBERT (editors), CRC press, January 2019, pp. 1-21, https://hal.archives-ouvertes.fr/hal-01963798

### Scientific Popularization

[62] V. MUÑOZ RAMÍREZ, F. FORBES, A. ARNAUD, E. MORO, M. DOJAT. *Anomaly detection in the MRI data of newly diagnosed Parkinsonian patients*, March 2019, 4e congrès de la Société Française de Résonance Magnétique en Biologie et Médecine - SFRMBM 2019, Poster, https://hal.inria.fr/hal-02436613

### Other Publications

[63] J. ARBEL, R. CORRADIN, B. NIPOTI. *Dirichlet process mixtures under affine transformations of the data*, January 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01950652

[64] J. ARBEL, O. MARCHAL, B. NIPOTI. *On the Hurwitz zeta function with an application to the exponential-beta distribution*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02400451

[65] A. ARNAUD, F. FORBES, R. STEELE, B. LEMASSON, E. L. BARBIER. *Bayesian mixtures of multiple scale distributions*, September 2019, working paper or preprint, https://hal.inria.fr/hal-01953393

[66] K. ASHURBEKOVA, A. USSEGLIO-CARLEVE, F. FORBES, S. ACHARD. *Optimal shrinkage for robust covariance matrix estimators in a small sample size setting*, November 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02378034

[67] M. BOUSEBATA, G. ENJOLRAS, S. GIRARD. *Bayesian estimation of natural extreme risk measures. Application to agricultural insurance*, June 2019, Global Challenges Science Week: International interdisciplinary days of Grenoble Alpes, Poster, https://hal.archives-ouvertes.fr/hal-02150604

[68] M. BOUSEBATA, S. GIRARD, G. ENJOLRAS. *Estimation bayésienne des mesures de risques naturels extrêmes. Application à l'assurance du risque agricole*, March 2019, 1 p. , Assises Nationales des Risques Naturels 2019, Poster, https://hal.archives-ouvertes.fr/hal-02092358

[69] F. BOUX, F. FORBES, J. ARBEL, E. L. BARBIER. *Inverse regression in MR Fingerprinting: reducing dictionary size while increasing parameters accuracy*, October 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02314026

[70] A. CONSTANTIN, M. FAUVEL, S. GIRARD, S. IOVLEFF. *Classification de Signaux Multidimensionnels Irrégulièrement Échantillonnés*, August 2019, GRETSI 2019 - 27e Colloque francophone de traitement du signal et des images, Poster, https://hal.archives-ouvertes.fr/hal-02276255

[71] A. CONSTANTIN, M. FAUVEL, S. GIRARD, S. IOVLEFF. *Supervised classification of multidimensional and irregularly sampled signals*, April 2019, 1 p. , Statlearn 2019 - Workshop on Challenging problems in Statistical Learning, Poster, https://hal.archives-ouvertes.fr/hal-02092347

[72] L. GARDES, S. GIRARD. *On the estimation of the variability in the distribution tail*, December 2019, working paper or preprint, https://hal.inria.fr/hal-02400320

[73] S. GIRARD. *Deux méthodes statistiques pour la classification et la régression en grande dimension*, June 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02156948

[74] S. GIRARD, G. STUPFLER, A. USSEGLIO-CARLEVE. *Nonparametric extreme conditional expectile estimation*, May 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02114255

[75] S. GIRARD, G. STUPFLER, A. USSEGLIO-CARLEVE. *An Lp-quantile methodology for tail index estimation*, January 2020, working paper or preprint, https://hal.inria.fr/hal-02311609

[76] B. KUGLER, F. FORBES, S. DOUTÉ. *Massive hyperspectral images analysis by inverse regression of physical models*, April 2019, StatLearn 2019 Workshop on Challenging problems in Statistical Learning, Poster, https://hal.archives-ouvertes.fr/hal-02423640

[77] H. LU, J. ARBEL, F. FORBES. *Bayesian nonparametric priors for hidden Markov random fields*, June 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02163046

[78] H. LU, F. FORBES, J. ARBEL. *Bayesian Nonparametric Priors for Graph Structured Data: Application to Image Segmentation*, January 2020, Bayes Comp 2020, Poster, https://hal.archives-ouvertes.fr/hal-02423642

[79] H. D. NGUYEN, J. ARBEL, H. LU, F. FORBES. *Approximate Bayesian computation via the energy statistic*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02399934

## References in notes

[80] C. BOUVEYRON. *Modélisation et classification des données de grande dimension. Application à l'analyse d'images*, Université Grenoble 1, septembre 2006, http://tel.archives-ouvertes.fr/tel-00109047

[81] P. EMBRECHTS, C. KLÜPPELBERG, T. MIKOSH. *Modelling Extremal Events*, Applications of Mathematics, Springer-Verlag, 1997, vol. 33

[82] F. FERRATY, P. VIEU. *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Springer, 2006

[83] S. GIRARD. *Construction et apprentissage statistique de modèles auto-associatifs non-linéaires. Application à l'identification d'objets déformables en radiographie. Modélisation et classification*, Université de Cery-Pontoise, octobre 1996

[84] K. LI. *Sliced inverse regression for dimension reduction*, in "Journal of the American Statistical Association", 1991, vol. 86, pp. 316–327

[85] B. OLIVIER, J.-B. DURAND, A. GUÉRIN-DUGUÉ, M. CLAUSEL. *Eye-tracking data analysis using hidden semi-Markovian models to identify and characterize reading strategies*, in "19th European Conference on Eye Movements (ECM 2017)", Wuppertal, Germany, August 2017, https://hal.inria.fr/hal-01671224

[86] J. SIMOLA, J. SALOJÄRVI, I. KOJO. *Using hidden Markov model to uncover processing states from eye movements in information search tasks*, in "Cognitive Systems Research", Oct 2008, vol. 9, n$^{o}$ 4, pp. 237-251