IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2014

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

Keywords: Stochastic Models, Machine Learning, Data Analysis, Image Processing, Statistical Methods

Creation of the Project-Team: 2008 January 01.

1. Members

Research Scientists
Florence Forbes [Team leader, Inria, Senior Researcher, HdR]
Stéphane Girard [Inria, Researcher, HdR]

Faculty Members
Jean-Baptiste Durand [Grenoble INP, Associate Professor]
Marie-José Martinez [Univ. Grenoble II, Associate Professor]

Engineers
Flor Vasseur [Inria, until Mar 2014]
Thomas Perret [Inria, from Sep 2014]

PhD Students
Alessandro Chiancone [Univ. Grenoble I, from Oct 2013]
Alexis Arnaud [Univ. Grenoble I, from Oct 2014]
Aina Frau Pascual [Inria, Cordi-S, from Oct 2013]
Gildas Mazo [Inria, Cordi-S, until Nov 2014]

Post-Doctoral Fellows
Thomas Vincent [Inria, until Apr 2014]
Angelika Studeny [Inria, until Jun 2014]
Farida Enikeeva [Ensimag, until Aug 2014]
Pablo Mesejo Santiago [Inria, from Sep 2014]

Visiting Scientists
Darren Wraith [QUT, Brisbane, Australia, Jul 2014]
Seydou-Nourou Sylla [IRD, PhD, from Sep 2014]

Administrative Assistants
Françoise de Coninck [Inria, until Jun 2014]
Giselle Loquet [Inria, from Sep 2014]

Others
Alexis Arnaud [Inria, Intern, from Feb 2014 until Jun 2014]
Anne Charlier [Inria, Intern, from Apr 2014 until Jul 2014]
Qianru Lisa [Hemera, Intern, from Apr 2014 until Jul 2014]

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, these 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 connected with 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: Angelika Studeny, Thomas Vincent, Alexis Arnaud, Jean-Baptiste Durand, Florence Forbes, Aina Frau Pascual, Alessandro Chiancone, Stéphane Girard, Marie-José Martinez.

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

In a first approach, we consider statistical parametric models, θ being the parameter, possibly multi-dimensional, usually unknown and to be estimated. We consider cases where the data naturally divides into observed data $y = y_1, ..., y_n$ 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,

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

These models are interesting in that they may point out hidden variable 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: Angelika Studeny, Thomas Vincent, Jean-Baptiste Durand, Florence Forbes.

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: Farida Enikeeva, Alessandro Chiancone, Stéphane Girard, Gildas Mazo, Seydou-Nourou Sylla, Pablo Mesejo Santiago.

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 [71]. 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 [74] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [69], 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 $X_{n,n}$ (see Figure 1).

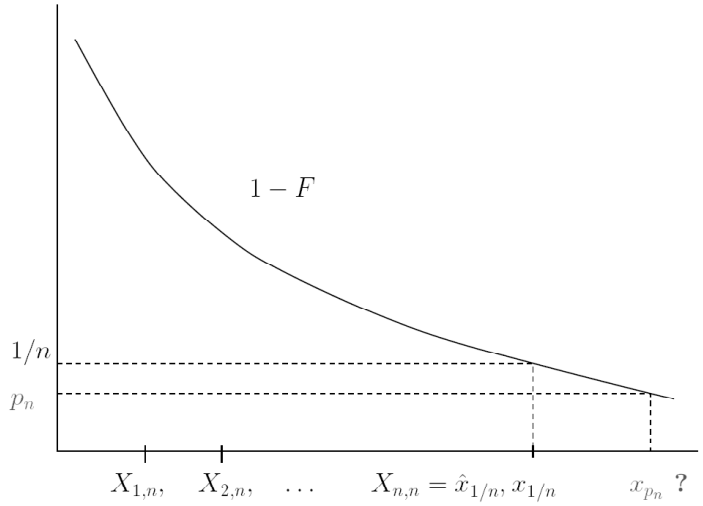

Figure 1. The curve represents the survival function $x \to P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\widehat{x}_{1/n} = X_{n,n}$. As illustrated in the figure, to estimate p_n-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.

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

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

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

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

for all $t > 0$. The function $\ell(x)$ acts as a nuisance parameter which yields a bias in the classical extreme-value estimators developed so far. Such models are often referred to as heavy-tail models since the probability of extreme events decreases at a polynomial rate to zero. It may be necessary to refine the model (2,3) by specifying a precise rate of convergence in (3). To this end, a second order condition is introduced involving an additional parameter $\rho \leq 0$. The larger ρ is, the slower the convergence in (3) and the more difficult the estimation of extreme quantiles.
More generally, the problems that we address are part of the risk management theory. For instance, in reliability, the distributions of interest are included in a semi-parametric family whose tails are decreasing exponentially fast. These so-called Weibull-tail distributions [9] 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 distorsions 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 [72]. 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 [67]. 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 [74].

4. Application Domains

4.1. Image Analysis

Participants: Alexis Arnaud, Aina Frau Pascual, Thomas Vincent, Florence Forbes, Stéphane Girard, Flor Vasseur, Alessandro Chiancone, Farida Enikeeva, Thomas Perret, Pablo Mesejo Santiago.

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 data, in collaboration with the Grenoble Institute of Neuroscience (GIN) and the NeuroSpin center of CEA Saclay. We also consider other statistical 2D fields coming from other domains such as remote sensing, in collaboration with Laboratoire de Planétologie de Grenoble. We worked on hyperspectral images. In the context of the "pole de competivite" project I-VP, we worked of images of PC Boards.

4.2. Biology, Environment and Medicine

Participants: Thomas Vincent, Aina Frau Pascual, Florence Forbes, Stéphane Girard, Gildas Mazo, Angelika Studeny, Seydou-Nourou Sylla, Marie-José Martinez, Jean-Baptiste Durand.

A second domain of applications concerns biology and medicine. We consider the use of missing data models in epidemiology. We also investigated statistical tools for the analysis of bacterial genomes beyond gene detection. Applications in neurosiences are also considered. Finally, in the context of the ANR VMC project Medup, we studied the uncertainties on the forecasting and climate projection for Mediterranean high-impact weather events.

5. New Software and Platforms

5.1. The LOCUS software

Participant: Florence Forbes.

Joint work with: Senan Doyle (start-up creator) and Michel Dojat from Grenoble Institute of Neuroscience and Benoit Scherrer from Harvard Medical School, Boston, MA, USA.

From brain MR images, neuroradiologists are able to delineate tissues such as grey matter and structures such as Thalamus and damaged regions. This delineation is a common task for an expert but unsupervised segmentation is difficult due to a number of artefacts. The LOCUS software (http://locus.gforge.inria.fr) automatically perform this segmentation for healthy brains An image is divided into cubes on each of which a statistical model is applied. This provides a number of local treatments that are then integrated to ensure consistency at a global level, resulting in low sensitivity to artifacts. The statistical model is based on a Markovian approach that enables to capture the relations between tissues and structures, to integrate a priori anatomical knowledge and to handle local estimations and spatial correlations.

The LOCUS software has been developed in the context of a collaboration between Mistis, a computer science team (Magma, LIG) and a Neuroscience methodological team (the Neuroimaging team from Grenoble Institut of Neurosciences, INSERM). This collaboration resulted over the period 2006-2008 into the PhD thesis of B. Scherrer (advised by C. Garbay and M. Dojat) and in a number of publications. In particular, B. Scherrer received a "Young Investigator Award" at the 2008 MICCAI conference.

The originality of this work comes from the successful combination of the teams respective strengths i.e. expertise in distributed computing, in neuroimaging data processing and in statistical methods.

5.2. The P-LOCUS software

Participants: Florence Forbes, Flor Vasseur.

Joint work with: Senan Doyle (start-up creator) and Michel Dojat.

The Locus software was extended to address the delineation of lesions in pathological brains. Its extension P-LOCUS (http://p-locus.com) for lesion detection was realized by S. Doyle with financial support from GRAVIT (Grenoble Alpes Valorisation Innovation Technologies, http://www.gravit-innovation.org/) with the goal to create a Start-up. P-LOCUS software analyses, in few minutes, a 3D MR brain scan and performs fully automatic brain lesion delineation using a combined dataset of various 3D MRI sequences. Its originality comes from:

- it is fully automatic: no external user interaction and no training data required
- the possibility to combine information from several images (MR sequences)
- a statistical Bayesian framework for robustness to image artefacts and a priori knowledge incorporation
- a voxel-based clustering technique that uses Markov random fields (MRF) incorporating information about neighboring voxels for spatial consistency and robustness to imperfect image features (noise).
- the possibility to select and incorporate relevant a priori knowledge via different atlases, e.g. tissue and vascular territory atlases
- a fully integrated preprocessing steps and lesion ROI identification

P-LOCUS software was presented at various conferences and used for the BRATS Challenge on tumor segmentation organized as a satellite challenge of the Miccai conference in Nagoya, Japan. A paper published in IEEE trans. on Medical Imaging reports the challenge results [24]. Results are also shown in [47]. The software has been registered at APP in 2013 and is now undergoing industrial development for the creation of a start-up (Pixyl) expected in January 2015.

5.3. The PyHRF software

Participants: Thomas Perret, Florence Forbes, Thomas Vincent, Aina Frau Pascual.

Joint work with: Philippe Ciuciu and Solveig Badillo from Parietal Team Inria and CEA NeuroSpin, Lotfi Chaari and Laurent Risser from INP Toulouse.

As part of fMRI data analysis, the PyHRF package (http://pyhrf.org) 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. With respect to the sole detection issue (i), the classical voxelwise GLM procedure is also available through NIPY, whereas Finite Impulse Response (FIR) and temporally regularized FIR models are implemented to deal with the HRF estimation concern (ii). Several parcellation tools are also integrated such as spatial and functional clusterings. Parcellations may be used for spatial averaging prior to FIR/RFIR analysis or to specify the spatial support of the HRF estimates in the JDE approach. These analysis procedures can be applied either to volumic data sets or to data projected onto the cortical surface. For validation purpose, this package is shipped with artificial and real fMRI data sets. To cope with the high computational needs for inference, PyHRF handles distributing computing by exploiting cluster units as well as multiple cores computers. Finally, a dedicated viewer is available which handles n-dimensional images and provides suitable features for exploring whole brain hemodynamics (display of time series, maps, ROI mask overlay). A paper in Frontiers in Neuroinformatics gives more details on the current PyHRF functionalities [26]. The 2-year engineer position of Thomas Perret is devoted to this software development.

5.4. R packages

Participants: Florence Forbes, Stéphane Girard, Gildas Mazo, Alexis Arnaud.

Joint work with: Charles Bouveyron (Univ. Paris 5) and Stéthane Dépréaux (LJK).

MISTIS is involved in the development of several R packages available on the CRAN archive. They are dedicated to the construction of copulas and to the classification and clustering of data.

- **PBC** (product of bivariate copulas). http://cran.r-project.org/web/packages/PBC/ This R package provides tools for building copulas with the PBC model, a class of multivariate copulas based on Products of Bivariate Copulas. Copulas are a useful tool to model multivariate distributions. While there exist various families of bivariate copulas, much fewer has been done when the dimension is higher. To this aim an interesting class of copulas based on products of transformed copulas has been proposed. However the use of this class for practical high dimensional problems remains challenging. Constraints on the parameters and the product form render inference, and in particular the likelihood computation, difficult. In this R package, we propose a new class of high dimensional copulas based on a product of transformed bivariate copulas. No constraints on the parameters refrain the applicability of the proposed class which is well suited for applications in high dimension. Furthermore the analytic forms of the copulas within this class allow to associate a natural graphical structure (see illustration below) which helps to visualize the dependencies and to compute the likelihood efficiently even in high dimension.
- **FDG** (one-Factor copulas with Durante Generators). http://cran.r-project.org/web/packages/FDGcopulas/ This R package provides tools for building high-dimensional copulas with the FDG model, a class of multivariate copulas based on one-factor copulas. FDG copulas are a class of copulas featuring an interesting balance between flexibility and tractability. This package provides tools to construct, calculate the pairwise dependence coefficients of, simulate from, and fit FDG copulas. The acronym FDG stands for 'one-Factor with Durante Generators', as an FDG copula is a one-factor copula - that is, the variables are independent given a latent factor - whose linking copulas belong to the Durante class of bivariate copulas (also referred to as exchangeable Marshall-Olkin or semilinear copulas).
- **HDclassif** (classification and clustering methods for high dimensional data). http://cran.r-project.org/web/packages/HDclassif/ The HDclassif package is devoted to the clustering and the discriminant analysis of high-dimensional data. The classification methods proposed in the package result from a new parametrization of the Gaussian mixture model which combines the idea of dimension

reduction and model constraints on the covariance matrices. The supervised classification method using this parametrization has been called High Dimensional Discriminant Analysis (HDDA). In a similar manner, the associated clustering method has been called High Dimensional Data Clustering (HDDC) and uses the Expectation-Maximization (EM) algorithm for inference. In order to correctly fit the data, both methods estimate the specific subspace and the intrinsic dimension of the groups. Due to the constraints on the covariance matrices, the number of parameters to estimate is significantly lower than other model-based methods and this allows the methods to be stable and efficient in high-dimensional spaces. Experiments on artificial and real datasets show that HDDC and HDDA perform better than existing classical methods on high-dimensional datasets, even with small datasets.

- **robustDA** (robust mixture discriminant analysis). http://cran.r-project.org/web/packages/robustDA/ Robust mixture discriminant analysis allows to build a robust supervised classifier from learning data with label noise. The idea of the proposed method is to confront an unsupervised modeling of the data with the supervised information carried by the labels of the learning data in order to detect inconsistencies. The method is able afterward to build a robust classifier taking into account the detected inconsistencies into the labels. An application to object recognition under weak supervision is presented below.
- **MSST** (Mixtures of multiple scaled Student distributions). The package is not yet available on the CRAN but should be early 2015. It implements more efficiently the models and inference procedures described in [21] and will be used on large data sets of brain MRI in the context of Alexis Arnaud PhD thesis. This is joint work with S. Dépréaux who helped with writing subroutines in C++.

6. New Results

6.1. Highlights of the Year

6.1.1. P-Locus software and Pixyl start-up project

The work on the P-Locus software has been exploited in order to create a start-up in January 2015. The project called Pixyl have been accepted by the GATE1 incubator and has been awarded a BPI emergence prize. It is leaded by Senan Doyle (future CEO). The other co-founders are Michel Dojat (INSERM, GIN), Florence Forbes (Inria, Mistis) and IT-Translation.

6.2. Mixture models

6.2.1. Parameter estimation in the heterogeneity linear mixed model

Participant: Marie-José Martinez.

Joint work with: Emma Holian (National University of Ireland, Galway)

In studies where subjects contribute more than one observation, such as in longitudinal studies, linear mixed models have become one of the most used techniques to take into account the correlation between these observations. By introducing random effects, mixed models allow the within-subject correlation and the variability of the response among the different subjects to be taken into account. However, such models are based on a normality assumption for the random effects and reflect the prior belief of homogeneity among all the subjects. To relax this strong assumption, Verbeke and Lesaffre (1996) proposed the extension of the classical linear mixed model by allowing the random effects to be sampled from a finite mixture of normal distributions with common covariance matrix. This extension naturally arises from the prior belief of the presence of unobserved heterogeneity in the random effects population. The model is therefore called the heterogeneity linear mixed model. Note that this model does not only extend the assumption about the random effects distribution, indeed, each component of the mixture can be considered as a cluster containing a proportion of the total population. Thus, this model is also suitable for classification purposes.

Concerning parameter estimation in the heterogeneity model, the use of the EM-algorithm, which takes into account the incomplete structure of the data, has been considered in the literature. Unfortunately, the M-step in the estimation process is not available in analytic form and a numerical maximisation procedure such as Newton-Raphson is needed. Because deriving such a procedure is a non-trivial task, Komarek et al. (2002) proposed an approximate optimization. But this procedure proved to be very slow and limited to small samples due to requiring manipulation of very large matrices and prohibitive computation.

To overcome this problem, we have proposed in an alternative approach which consists of fitting directly an equivalent mixture of linear mixed models. Contrary to the heterogeneity model, the M-step of the EM-algorithm is tractable analytically in this case. Then, from the obtained parameter estimates, we can easily obtain the parameter estimates in the heterogeneity model.

6.2.2. Taking into account the curse of dimensionality

Participants: Stéphane Girard, Alessandro Chiancone, Seydou-Nourou Sylla.

Joint work with: C. Bouveyron (Univ. Paris 5), M. Fauvel (ENSAT Toulouse) and J. Chanussot (Gipsa-lab and Grenoble-INP)

In the PhD work of Charles Bouveyron (co-advised by Cordelia Schmid from the Inria LEAR team) [67], we propose new Gaussian models of high dimensional data for classification purposes. We assume that the data live in several groups located in subspaces of lower dimensions. Two different strategies arise:

- the introduction in the model of a dimension reduction constraint for each group
- the use of parsimonious models obtained by imposing to different groups to share the same values of some parameters

This modelling yields a new supervised classification method called High Dimensional Discriminant Analysis (HDDA) [4]. Some versions of this method have been tested on the supervised classification of objects in images. This approach has been adapted to the unsupervised classification framework, and the related method is named High Dimensional Data Clustering (HDDC) [3]. Our recent work consists in adding a kernel in the previous methods to deal with nonlinear data classification and heterogeneous data [12]. We also investigate the use of kernels derived from similary measures on binary data. The targeted application is the analysis of verbal autopsy data (PhD thesis of N. Sylla): Indeed, health monitoring and evaluation make more and more use of data on causes of death from verbal autopsies in countries which do not keep records of civil status or with incomplete records. The application of verbal autopsy method allows to discover probable cause of death. Verbal autopsy has become the main source of information on causes of death in these populations.

6.2.3. Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering

Participant: Florence Forbes.

Joint work with: Darren Wraith from QUT, Brisbane Australia.

Clustering concerns the assignment of each of N, possibly multidimensional, observations $y_1, ..., y_N$ to one of K groups. A popular way to approach this task is via a parametric finite mixture model. While the vast majority of the work on such mixtures has been based on Gaussian mixture models in many applications the tails of normal distributions are shorter than appropriate or parameter estimations are affected by atypical observations (outliers). The family of location and scale mixtures of Gaussians has the ability to generate a number of flexible distributional forms. It nests as particular cases several important asymmetric distributions like the Generalised Hyperbolic (GH) distribution. The Generalised Hyperbolic distribution in turn nests many other well known distributions such as the Normal Inverse Gaussian (NIG) whose practical relevance has been widely documented in the literature. In a multivariate setting, we propose to extend the standard location and scale mixture concept into a so called multiple scaled framework which has the advantage of allowing different tail and skewness behaviours in each dimension of the variable space with arbitrary correlation between dimensions. The approach builds upon, and develops further, previous work on scale mixtures of

Gaussians [21]. Estimation of the parameters is provided via an EM algorithm with a particular focus on NIG distributions. Inference is then extended to cover the case of mixtures of such multiple scaled distributions for application to clustering. Assessments on simulated and real data confirm the gain in degrees of freedom and flexibility in modelling data of varying tail behaviour and directional shape. In addition, comparison with other similar models of GH distributions shows that the later are not as flexible as claimed.

6.2.4. Bayesian mixtures of multiple scaled distributions

Participants: Florence Forbes, Alexis Arnaud.

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

In previous work [21], inference for mixtures of multiple scaled distributions has been carried out based on maximum likelihood principle and using the EM algorithm. In this work we consider a Bayesian treatment of these models for the many advantages that the Bayesian framework offers in the mixture model context. Mainly it avoids the ill-posed nature of maximum likelihood due to the presence of singularities in the likelihood function. A mixture component may collapse by becoming centered at a single data vector sending its covariance to 0 and the model likelihood to infinity. A Bayesian treatment protects the algorithm from this problem occurring in ordinary EM. Also, Bayesian model comparison embodies the principle that states that simple models should be preferred. Typically, maximum likelihood does not provide any guidance on the choice of the model order as more complex models can always fit the data better. For standard scale mixture of Gaussians, the usual Normal-Wishart prior can be used for the Gaussian parameters. For multiple scaled distributions, the specific decomposition of the covariance requires appropriate separated priors on the eigenvectors and eigenvalues of the scale matrix. Such a decomposition has been already examined in various works on priors for covariance matrix. In this work we consider several possibilities. We derive an inference scheme based on variational approximation and show how to apply this to model selection. In particular, we consider the issue of selecting automatically an appropriate number of classes in the mixtures. We show how to select this number from a single run avoiding the repetitive inference and comparison of all possible models.

6.2.5. EM for Weighted-Data Clustering

Participant: Florence Forbes.

Joint work with: Israel Gebru, Xavier Alameda-Pined and Radu Horaud from the Inria Perception team.

Data clustering has received a lot of attention and many methods, algorithms and software packages are currently available. Among these techniques, parametric finite-mixture models play a central role due to their interesting mathematical properties and to the existence of maximum-likelihood estimators based on expectation-maximization (EM). In this work we propose a new mixture model that associates a weight with each observed data point. We introduce a Gaussian mixture with weighted data and we derive two EM algorithms: the first one assigns a fixed weight to each observed datum, while the second one treats the weights as hidden variables drawn from gamma distributions. We provide a general-purpose scheme for weight initialization and we thoroughly validate the proposed algorithms by comparing them with several parametric and non-parametric clustering techniques. We demonstrate the utility of our method for clustering heterogeneous data, namely data gathered with different sensorial modalities, e.g., audio and vision. See also an application in [40].

6.3. Statistical models for Neuroscience

6.3.1. Physiologically informed Bayesian analysis of ASL fMRI data

Participants: Florence Forbes, Aina Frau Pascual, Thomas Vincent.

Joint work with: Philippe Ciuciu from Team Parietal and Neurospin, CEA in Saclay.

ASL fMRI data provides a quantitative measure of blood perfusion, that can be correlated to neuronal activation. In contrast to BOLD measure, it is a direct measure of cerebral blood flow. However, ASL data has a lower SNR and resolution so that the recovery of the perfusion response of interest suffers from the contamination by a stronger BOLD component in the ASL signal. In this work [38], [39] we consider a model of both BOLD and perfusion components within the ASL signal. A physiological link between these two components is analyzed and used for a more accurate estimation of the perfusion response function in particular in the usual ASL low SNR conditions.

6.3.2. Physiological models comparison for the analysis of ASL fMRI data

Participants: Florence Forbes, Aina Frau Pascual.

Joint work with: Philippe Ciuciu from Team Parietal and Neurospin, CEA in Saclay.

Physiological models have been proposed to describe the processes that underlie the link between neural and hemodynamic activity in the brain. Among these, the Balloon model describes the changes in blood flow, blood volume and oxygen concentration when an hemodynamic response is ensuing neural activation. Next, a *BOLD signal model* links these variables to the measured BOLD signal. Taken together, these equations allow the precise modeling of the coupling between the cerebral blood flow (CBF) and hemodynamic response (HRF). However, several competing versions of BOLD signal model have been described in the past. In this work, we compare different physiological models linking CBF to HRF and different BOLD signal models too in terms of least squares error and log-likelihood, and we assess the impact of this setting in the context of Arterial Spin Labelling (ASL) functional Magnetic Resonance Imaging (fMRI) data analysis.

6.3.3. Variational EM for the analysis of ASL fMRI data

Participants: Florence Forbes, Aina Frau Pascual.

Joint work with: Philippe Ciuciu from Team Parietal and Neurospin, CEA in Saclay.

In this work, the goal is to analyse ASL data by accounting jointly for both the BOLD and perfusion components in the signal. Using the model proposed in [77], we design a variational EM approach to estimate the model parameters as a faster alternative to the MCMC approach used in [77] and [39].

6.3.4. Metaheuristics for the analysis of fMRI data

Participants: Florence Forbes, Pablo Mesejo Santiago.

Joint work with: Jan Warnking from Grenoble Institute of Neuroscience.

The undergoing work is focused on the optimization of nonlinear models for fMRI data analysis, specially Blood-oxygen-level dependent (BOLD) MR modality. The current optimization procedure consists of a Bayesian inversion of the nonlinear model using a Gauss-Newton/Expectation-Maximization algorithm. Such an optimization procedure is time-consuming and achieves sub-optimal results. Therefore, the current research work is mainly focused on improving these results by experimenting with global search optimization methods, like metaheuristics (MHs). Secondly, MHs can also be of great help in the development of minimization algorithms for solving problems with orthogonality constraints (like in polynomial optimization, combinatorial optimization, eigenvalue problems, sparse PCA, matrix rank minimization, etc.). Thus, another main research line is concerned with the application of MHs to this problem and, if necessary, the design and implementation of new evolutionary operators that preserve orthogonality. And, finally, we are also trying to create advanced statistical models for coupling Arterial Spin Labeling (ASL) and BOLD MR modalities to study brain function.

6.3.5. Model selection for hemodynamic brain parcellation in fMRI

Participant: Florence Forbes.

Joint work with: Lotfi Chaari, Mohanad Albughdadi, Jean-Yves Tourneret from IRIT-ENSEEIHT in Toulouse and Philippe Ciuciu from Neurospin, CEA in Saclay.

Brain parcellation into a number of hemodynamically homogeneous regions (parcels) is a challenging issue in fMRI analyses. This task has been recently integrated in the joint detection-estimation (JDE) resulting in the so-called joint detection-parcellation-estimation (JPDE) model. JPDE automatically estimates the parcels from the fMRI data but requires the desired number of parcels to be fixed. This is potentially critical in that the chosen number of parcels may influence detection-estimation performance. In this paper [30], we propose a model selection procedure to automatically fix the number of parcels from the data. The selection procedure relies on the calculation of the free energy corresponding to each concurrent model, within the variational expectation maximization framework. Experiments on synthetic and real fMRI data demonstrate the ability of the proposed procedure to select an adequate number of parcels. We also investigated the use of Latent Dirichlet Processes.

6.3.6. Partial volume estimation in brain MRI revisited

Participant: Florence Forbes.

Joint work with: Alexis Roche from Siemens Advanced Clinical Imaging Technology, Department of Radiology, CHUV, Signal Processing Laboratory (LTS5), EPFL, Lausanne, Switzerland.

Image-guided diagnosis of brain disease calls for accurate morphometry algorithms, e.g., in order to detect focal atrophy patterns relating to early-stage progression of particular forms of dementia. To date, widely used brain morphometry packages rest upon discrete Markov random field (MRF) image segmentation models that ignore, or do not fully account for partial voluming, leading to potentially inaccurate estimation of tissue volumes. Although several partial volume (PV) estimation methods have been proposed in the literature from the early 90's, none of them seems to be in common use. In [43], we propose a fast algorithm to estimate brain tissue concentrations from conventional T1-weighted images based on a Bayesian maximum a posteriori formulation that extends the "mixel" model developed in the 90's. A key observation is the necessity to incorporate additional prior constraints to the "mixel" model for the estimation of plausible concentration maps. Experiments on the ADNI standardized dataset show that global and local brain atrophy measures from the proposed algorithm yield enhanced diagnosis testing value than with several widely used soft tissue labeling methods.

6.3.7. Tumor classification and prediction using robust multivariate clustering of multiparametric MRI

Participants: Florence Forbes, Alexis Arnaud.

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

Advanced statistical clustering approaches are promising tools to better exploit the wealth of MRI information especially on large cohorts and multi-center studies. In neuro-oncology, the use of multiparametric MRI may better characterize brain tumor heterogeneity. To fully exploit multiparametric MRI (e.g. tumor classification), appropriate analysis methods are yet to be developed. They offer improved data quality control by allowing automatic outlier detection and improved analysis by identifying discriminative tumor signatures with measurable predictive power. In this work, we show on small animals data that advanced statistical learning approaches can help 1) in organizing existing data by detecting and excluding outliers and 2) in building a dictionary of tumor fingerprints from a clustering analysis of their microvascular features. Future work should include the integration in a joint statistical model of both automatic ROI delineation and clustering for whole brain data analysis, with a better use of anatomical information. This work has been submitted to the ISMRM 2015 conference and accepted in the SFMRMB 2015 conference [45].

6.4. Markov models

6.4.1. Identifying Interactions between Tropical Plant Species: A Correlation Analysis of High-Throughput Environmental DNA Sequence Data based on Random Matrix Theory

Participants: Florence Forbes, Angelika Studeny.

This is joint work with: Eric Coissac and Pierre Taberlet from LECA (Laboratoire d'Ecologie Alpine) and Alain Viari from Inria team Bamboo.

The study of species cooccurence pattern has always been central to community ecology. The rise of high-throughput molecular methods and their use in ecology nowadays allows for a facilitated access to new data of an unprecedented quantity. We address the question about the identification of genuine species interactions in the light of these novel data. The statistical analysis has to be tailored to the data specifics: the large amount of available data as well as biases inherent to the data extraction methods. The latter can cause spurious interactions while the former complicates any statistical modelling approach. In addition, the resolution of the data provided is rarely on the species level. In this work, we conduct a thorough correlation analysis between MOTUs (molecular operating taxonomic unit) on different spatial scales to investigate global as well as local spatial pattern. Although this type of analysis is per se exploratory, we suggest it here in order to separate true species interaction from random pattern and to identify species subgroups for further in detail modelling. A random-matrix approach allows us to derive objective cut-off values for genuine correlations. We compare the results with those derived by the application of a model-based, sparse regression approach. Our study shows that despite their seemingly less precise nature when it comes to species identification, these data enable us to reveal mechanisms that structure an ecological community. In the light of the nowadays facilitated access to molecular data, this points the way to a novel set of efficient methods for community analysis.

6.4.2. Modelling multivariate counts with graphical Markov models.

Participant: Jean-Baptiste Durand.

Joint work with: Pierre Fernique (Montpellier 2 University, CIRAD and Inria Virtual Plants) and Yann Guédon (CIRAD and Inria Virtual Plants)

Multivariate count data are defined as the number of items in different states issued from sampling within a population, which individuals own items in various numbers and states. The analysis of multivariate count data is a recurrent and crucial issue in numerous modelling problems, particularly in the fields of biology and ecology (where the data can represent, for example, children counts associated with multitype branching processes), sociology and econometrics. Denoting by K the number of states, multivariate count data analysis relies on modelling the joint distribution of the K-dimensional random vector $N = (N_0, ..., N_{K-1})$ with discrete components. Our work focused on I) Identifying states that appear simultaneously, or on the contrary that are mutually exclusive. This was achieved by identifying conditional independence relationships between the K variables; II)Building parsimonious parametric models consistent with these relationships; III) Characterizing and testing the effects of covariates on the distribution of N, particularly on the dependencies between its components.

Our context of application was characterised by zero-inflated, often right skewed marginal distributions. Thus, Gaussian and Poisson distributions were not *a priori* appropriate. Moreover, the multivariate histograms typically had many cells, most of which were empty. Consequently, nonparametric estimation was not efficient.

We developed an approach based on probabilistic graphical models (Koller & Friedman, 2009 [73]) to identify and exploit properties of conditional independence between numbers of children in different states, so as to simplify the specification of their joint distribution. The considered models are based on chain graphs. Model selection procedures are necessary to infer the graph and specify parsimonious distributions. The graph building stage was based on exploring the space of possible chain graph models, which required defining a notion of neighbourhood of these graphs. A parametric distribution was associated with each graph. It was obtained by combining families of univariate and multivariate distributions or regression models. These families were chosen by selection model procedures among different parametric families [36]. To relax the strong constraints regarding dependencies induced by using parametric distributions, mixture of graphical models were also considered [49].

Further extensions will be considered, and particularly

- Hidden Markov tree models (see 6.4.3) where the hidden state process is a multitype branching process with graphical generation distributions.

- Gaussian chain graph models, where the chain components can be identified using lasso methods.

6.4.3. Statistical characterization of tree structures based on Markov tree models and multitype branching processes, with applications to tree growth modelling.

Participant: Jean-Baptiste Durand.

Joint work with: Pierre Fernique (Montpellier 2 University and CIRAD) and Yann Guédon (CIRAD), Inria Virtual Plants.

Algorithmic issues in hidden Markov tree models were considered by Durand *et al.* (2004) [68]. This family of models was used to represent local dependencies and heterogeneity within tree-structured data. It relied on a tree-structured hidden state process, where the children states were assumed independent given their parent state. The latter assumption has been relaxed in an extension of these models and new algorithmic solutions for model inference have been proposed in Pierre Fernique's PhD [70]. An application to the study of the cell lineage in biological tissues responsible for the plant growth has been considered. In this setting, the number of children is small (between 0 and 2) and a saturated model has been considered to model transitions between parent and configurations of children states. Extensions will be proposed, based on the parametric discrete multivariate distributions developed in Section 6.4.2.

6.4.4. Change-point models for tree-structured data

Participant: Jean-Baptiste Durand.

Joint work with: Pierre Fernique (Montpellier 2 University and CIRAD) and Yann Guédon (CIRAD), Inria Virtual Plants.

As an alternative to the hidden Markov tree models discussed in Section 6.4.3, subtrees with similar attributes can be identified using multiple change-point models. These approaches are well-developed in the context of sequence analysis, but their extensions to tree-structured data are not straightforward. Their advantage on hidden Markov models is to relax the strong constraints regarding dependencies induced by parametric distributions and local parent-children dependencies. Heuristic approaches for change-point detection in trees were proposed and applied to the analysis of patchiness patterns (consisting of canopies made of clumps of either vegetative or flowering botanical units) in mango trees [70].

6.4.5. Hidden Markov models for the analysis of eye movements

Participant: Jean-Baptiste Durand.

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 make some kinds of press reviews. 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 steps in the cognitive processes leading to decision. In HMMs, each phase is associated with a state of the Markov chain. The states are observed indirectly through eye-movements. Our approach was inspired by Simola *et al.* (2008) [76], but we used hidden semi-Markov models for better characterization of phase length distributions. The estimated HMM highlighted contrasted reading strategies (i.e., state transitions), with both individual and document-related variability.

However, the characteristics of eye movements within each phase tended to be poorly discriminated. As a result, high uncertainty in the phase changes arose, and it could be difficult to relate phases to known patterns in EEGs.

As a perspective, we aim at developing an integrated model coupling EEG and eye movements within one single HMM for better identification of the phases. Here, the coupling should incorporate some delay between the transitions in both (EEG and eye-movement) chains, since EEG patterns associated to cognitive processes occur lately with respect to eye-movement phases. Moreover, EEGs and scanpaths were recorded with different time resolutions, so that some resampling scheme must be added into the model, for the sake of synchronizing both processes. Probabilistic graphical models (see Section 6.4.2) will be inferred from the channel correlations to represent interactions between brain zones. The variability of these graphs is partly explained by individual differences in text exploration, which will have to be quantified.

6.4.6. *Hyper-Spectral Image Analysis with Partially-Latent Regression and Spatial Markov Dependencies*

Participant: Florence Forbes.

Joint work with: Antoine Deleforge, Sileye Ba and Radu Horaud from the Inria Perception team.

Hyper-spectral data can be analyzed to recover physical properties at large planetary scales. This involves resolving inverse problems which can be addressed within machine learning, with the advantage that, once a relationship between physical parameters and spectra has been established in a data-driven fashion, the learned relationship can be used to estimate physical parameters for new hyper-spectral observations. Within this framework, we propose a spatially-constrained and partially-latent regression method which maps high-dimensional inputs (hyper-spectral images) onto low-dimensional responses (physical parameters). The proposed regression model comprises two key features. Firstly, it combines a Gaussian mixture of locally-linear mappings (GLLiM) with a partially-latent response model described in [17]. While the former makes high-dimensional regression tractable, the latter enables to deal with physical parameters that cannot be observed or, more generally, with data contaminated by experimental artifacts that cannot be explained with noise models. Secondly, spatial constraints are introduced in the model through a Markov random field (MRF) prior which provides a spatial structure to the Gaussian-mixture hidden variables. Experiments conducted on a database composed of remotely sensed observations collected from the Mars planet by the Mars Express orbiter demonstrate the effectiveness of the proposed model. A preliminary version of the work can be found in [31].

6.5. Semi and non-parametric methods

6.5.1. *Conditional extremal events*

Participant: Stéphane Girard.

Joint work with: L. Gardes (Univ. Strasbourg), A. Daouia (Univ. Toulouse I and Univ. Catholique de Louvain), J. Elmethni (Univ. Paris 5) and S. Louhichi (Univ. Grenoble 1)

The goal of the PhD thesis of Alexandre Lekina was 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, the tail heaviness of Y depends on X, and thus the tail index as well as the extreme quantiles are also functions of the covariate. We combine nonparametric smoothing techniques [71] with extreme-value methods in order to obtain efficient estimators of the conditional tail index and conditional extreme quantiles. The strong consistency of such estimator is established in [53]. When the covariate is functional and random (random design) we focus on kernel methods [58].

Conditional extremes are studied in climatology where one is interested in how climate change over years might affect extreme temperatures or rainfalls. In this case, the covariate is univariate (time). Bivariate examples include the study of extreme rainfalls as a function of the geographical location. The application part of the study is joint work with the LTHE (Laboratoire d'étude des Transferts en Hydrologie et Environnement) located in Grenoble.

6.5.2. Estimation of extreme risk measures

Participant: Stéphane Girard.

Joint work with: E. Deme (Univ. Gaston-Berger, Sénégal, J. Elmethni (Univ. Paris 5), L. Gardes and A. Guillou (Univ. Strasbourg)

One of the most popular risk measures is the Value-at-Risk (VaR) introduced in the 1990's. In statistical terms, the VaR at level $\alpha \in (0,1)$ corresponds to the upper α-quantile of the loss distribution. The Value-at-Risk however suffers from several weaknesses. First, it provides us only with a pointwise information: VaR(α) does not take into consideration what the loss will be beyond this quantile. Second, random loss variables with light-tailed distributions or heavy-tailed distributions may have the same Value-at-Risk . Finally, Value-at-Risk is not a coherent risk measure since it is not subadditive in general. A coherent alternative risk measure is the Conditional Tail Expectation (CTE), also known as Tail-Value-at-Risk, Tail Conditional Expectation or Expected Shortfall in case of a continuous loss distribution. The CTE is defined as the expected loss given that the loss lies above the upper α-quantile of the loss distribution. This risk measure thus takes into account the whole information contained in the upper tail of the distribution. It is frequently encountered in financial investment or in the insurance industry. In [52], we have established the asymptotic properties of the CTE estimator in case of extreme losses, *i.e.* when $\alpha \to 0$ as the sample size increases. We have exhibited the asymptotic bias of this estimator, and proposed a bias correction based on extreme-value techniques. In [20], we study the situation where some covariate information is available. We thus has to deal with conditional extremes (see paragraph 6.5.1). We also proposed a new risk measure (called the Conditional Tail Moment) which encompasses various risk measures, such as the CTE, as particular cases.

6.5.3. Multivariate extremal events

Participants: Stéphane Girard, Gildas Mazo, Florence Forbes.

Joint work with: C. Amblard (TimB in TIMC laboratory, Univ. Grenoble I), L. Gardes (Univ. Strasbourg) and L. Menneteau (Univ. Montpellier II)

Copulas are a useful tool to model multivariate distributions [75]. At first, we developed an extension of some particular copulas [1]. It followed a new class of bivariate copulas defined on matrices [55] and some analogies have been shown between matrix and copula properties.

However, while there exist various families of bivariate copulas, much fewer has been done when the dimension is higher. To this aim an interesting class of copulas based on products of transformed copulas has been proposed in the literature. The use of this class for practical high dimensional problems remains challenging. Constraints on the parameters and the product form render inference, and in particular the likelihood computation, difficult. We proposed a new class of high dimensional copulas based on a product of transformed bivariate copulas [64]. No constraints on the parameters refrain the applicability of the proposed class which is well suited for applications in high dimension. Furthermore the analytic forms of the copulas within this class allow to associate a natural graphical structure which helps to visualize the dependencies and to compute the likelihood efficiently even in high dimension. The extreme properties of the copulas are also derived and an R package has been developed.

As an alternative, we also proposed a new class of copulas constructed by introducing a latent factor. Conditional independence with respect to this factor and the use of a nonparametric class of bivariate copulas lead to interesting properties like explicitness, flexibility and parsimony. In particular, various tail behaviours are exhibited, making possible the modeling of various extreme situations [42]. A pairwise moment-based inference procedure has also been proposed and the asymptotic normality of the corresponding estimator has been established [66].

In collaboration with L. Gardes, we investigate the estimation of the tail copula which is widely used to describe the amount of extremal dependence of a multivariate distribution. In some situations such as risk management, the dependence structure can be linked with some covariate. The tail copula thus depends on this covariate and is referred to as the conditional tail copula. The aim of our work is to propose a nonparametric estimator of the conditional tail copula and to establish its asymptotic normality [57].

6.5.4. Level sets estimation

Participant: Stéphane Girard.

Joint work with: A. Guillou and L. Gardes (Univ. Strasbourg), A. Nazin (Univ. Moscou), G. Stupfler (Univ. Aix-Marseille) and A. Daouia (Univ. Toulouse I and Univ. Catholique de Louvain)

The boundary bounding the set of points is viewed as the larger level set of the points distribution. This is then an extreme quantile curve estimation problem. We proposed estimators based on projection as well as on kernel regression methods applied on the extreme values set, for particular set of points [10]. We also investigate the asymptotic properties of existing estimators when used in extreme situations. For instance, we have established in collaboration with G. Stupfler that the so-called geometric quantiles have very counter-intuitive properties in such situations [63], [62] and thus should not be used to detect outliers. These resuls are submitted for publication.
In collaboration with A. Daouia, we investigate the application of such methods in econometrics [16]: A new characterization of partial boundaries of a free disposal multivariate support is introduced by making use of large quantiles of a simple transformation of the underlying multivariate distribution. Pointwise empirical and smoothed estimators of the full and partial support curves are built as extreme sample and smoothed quantiles. The extreme-value theory holds then automatically for the empirical frontiers and we show that some fundamental properties of extreme order statistics carry over to Nadaraya's estimates of upper quantile-based frontiers.
In collaboration with A. Nazin, we define new estimators of the frontier function based on linear programming methods. The frontier is defined as the solution of a linear optimization problem under inequality constraints. The estimator is shown to be strongly consistent with respect to the L_1 norm and we establish that it reaches the optimal minimax rate of convergence [58].
In collaboration with G. Stupfler and A. Guillou, new estimators of the boundary are introduced. The regression is performed on the whole set of points, the selection of the "highest" points being automatically performed by the introduction of high order moments [22].

6.5.5. Retrieval of Mars surface physical properties from OMEGA hyperspectral images.

Participants: Stéphane Girard, Alessandro Chiancone.

Joint work with: S. Douté from Laboratoire de Planétologie de Grenoble, J. Chanussot (Gipsa-lab and Grenoble-INP) and J. Saracco (Univ. Bordeaux).

Visible and near infrared imaging spectroscopy is one of the key techniques to detect, to map and to characterize mineral and volatile (eg. water-ice) species existing at the surface of planets. Indeed the chemical composition, granularity, texture, physical state, etc. of the materials determine the existence and morphology of the absorption bands. The resulting spectra contain therefore very useful information. Current imaging spectrometers provide data organized as three dimensional hyperspectral images: two spatial dimensions and one spectral dimension. Our goal is to estimate the functional relationship F between some observed spectra and some physical parameters. To this end, a database of synthetic spectra is generated by a physical radiative transfer model and used to estimate F. The high dimension of spectra is reduced by Gaussian regularized sliced inverse regression (GRSIR) to overcome the curse of dimensionality and consequently the sensitivity of the inversion to noise (ill-conditioned problems) [15]. We have also defined an adaptive version of the method which is able to deal with block-wise evolving data streams [13].
In his PhD thesis work, Alessandro Chiancone studies the extension of the SIR method to different sub-populations. The idea is to assume that the dimension reduction subspace may not be the same for different clusters of the data [46]. He also published a paper on a previous work in the field of hierarchical segmentation of images [14].

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

A contract with the HEMERA company was contracted including the internships of Anne Charlier and Lisa Qianru. Hemera designs, produces and sells online liquid and gaz analyzers. It is located in Grenoble. The aim of Hemera is to measure, in any gaseous or liquid environment, with a minimalized environmental impact and in a selective way, all compounds seen nowadays as pollutants : for our health, for an industrial process, etc. Hemera's analyzers measure gaz concentrations using optical techniques. The goal of the collaboration was to investigate the use of statistical methods to improve both the determination of the present gaz and their respective concentrations from the analysis of spectra representing a mixture of the different gaz. A preliminary study based on the Lasso technique was implemented and tested with promising first conclusions.

8. Partnerships and Cooperations

8.1. Regional Initiatives

- **PERSYVACT project.** MISTIS is involved in a 2-year exploratory project, funded (20 keuros for the whole project) by the PERSYVAL labex (https://persyval-lab.org/en), with other teams from local laboratories, LJK, GIPSA-Lab and TIMC. The goal of this research project is to build tools for analyzing hierarchically structured models for high dimensional complex data. In parallel, MISTIS received **15 keuros** from the labex for the PhD of A. Chiancone co-advised with J. Chanussot from GIPSA-Lab.
- **Grenoble Pole Cognition (2013-14).** We received in 2012, 2013 and 2014 **2.5 keuros** from the Grenoble Pole Cogntion, http://www.grenoblecognition.fr/, for collaborative projects involving the GIN and NeuroSpin. This funding was used this year for the internship of Alexis Arnaud on MRI analysis for small animals.
- MISTIS is involved in three regional initiatives: PEPS (funded by CNRS and the PRES of Grenoble), AGIR (funded by Université Grenoble 1 and Grenoble-INP) and the MOTU project (funded by UPMF). The first two projects focus on the modelling of the extreme risk and its application in social science. The partners include the LTHE (Laboratoire d'étude des Transferts en Hydrologie et Environnement) and the 3S-R lab (Sols, Solides, Structures - Risques). The third project focuses on the use of statistical techniques for transportation data analysis and involves the GAEL laboratory (Grenoble Applied Economics Laboratory).
- MISTIS participates in the weekly statistical seminar of Grenoble. Jean-Baptiste Durand is in charge of the organization and several lecturers have been invited in this context.
- S. Girard is at the head of the probability and statistics department of the LJK since september 2012.

8.2. International Initiatives

8.2.1. 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 Washington in Seattle, the Russian Academy of Science in Moscow, the National University of Ireland in Galway, and more recent partners like IDIAP involved in the HUMAVIPS project, Université Gaston Berger in Senegal and University of Melbourne in Australia. We will also work at turning other current European contacts, *e.g.* at EPFL (A. Roche at University Hospital Lausanne and Siemens Healthcare), into more formal partnerships and eventually explore the possibility for a H2020 project in the *Personalizing Health and Care* axis.

The main international collaborations that we are currently trying to develop are with:

- Fabrizio Durante, Free University of Bozen-Bolzano, Italy.
- Emma Holian and John Hinde from National University of Ireland, Galway, Ireland.
- K. Qin and D. Wraith from RMIT in Melbourne, Australia and Queensland University of Technology in Brisbane, Australia.
- E. Deme and S. Sylla from Saint Louis university and IRD in Saint Louis, Senegal.
- Alexandre Nazin and Russian Academy of Science in Moscow, Russia.
- Alexis Roche and University Hospital Lausanne/Siemens Healthcare, Advanced Clinical Imaging Technology group, Lausanne, Switzerland.

8.3. International Research Visitors

8.3.1. Visits of International Scientists

- Seydou Nourou Sylla (Université Gaston Berger, Sénégal) has been hosted by the MISTIS team for four months.
- Darren Wraith (Queensland University of Technology in Brisbane, Australia) has been hosted by the MISTIS team for 2 weeks.

8.3.1.1. Internships

Alexis Arnaud (Master, from Feb 2014 until June 2013)

Subject: Mixtures of generalized Student multivariate distributions: application to tumor characterisation from multiparametric MRI.

Institution: University Montpellier 2

Anne Charlier (2nd year)

Subject: Estimation of gaz concentrations in a gaz mixture from spectrophotometric measures.

Institution: PHELMA, Grenoble-INP

Lisa Qian-ru (Master)

Subject: Inverse regression to identify and quantify polluants from UV spectroscopy measures.

Institution: Univ. PMF, Hemera, Meylan

Seydou-Nourou Sylla (PhD, from September 2014 to December 2014)

Subject: Classification for medical data

Institution: Université Gaston Berger (Sénégal)

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. general chair, scientific chair

- F. Forbes co-organized the workshop *Statistical Challenges in Neuroscience* in Warwick, UK in Sept. 2014, http://www2.warwick.ac.uk/fac/sci/statistics/crism/workshops/neuroscience/.
- F. Forbes co-organized the workshop on *Probabilisitic graphical models and structured data on graphs* in Grenoble, in July 2014.

- Stéphane Girard co-organized the workshop "Extreme Value Theory, Spatial and Temporal Aspects", Besançon, https://trimestres-lmb.univ-fcomte.fr/Workshop-on-Extreme-Value-Theory
- Stéphane Girard co-organized the "Rencontres d'Astrostatistique", Grenoble, http://astrostat2014.sciencesconf.org
- "Extremes and Copulas", Grenoble, http://mistis.inrialpes.fr/workshop-copulas-extremes.
- Stéphane Girard organized the workshop "*Copulas and extremes*", Grenoble, http://mistis.inrialpes.fr//workshop-copulas-extremes.html.
- Marie José Martinez, Jean Baptiste Durand, Florence Forbes in collaboration with Iragael Joly (Grenoble Applied Economics Laboratory) organized the workshop "Statistics, Activities and Transportation" in Grenoble http://mistis.inrialpes.fr/workshop-statistique-transport.html as part of the MOTU project (2013-14).

9.1.1.2. member of the organizing committee

- F. Forbes is a member of the committee for the 2nd SFRMBM (Société Francaise de Résonance Magnétique en Biologie et Médecine) conference in Grenoble in 2015, http://sfrmbm2015.sciencesconf.org/

9.1.1.3. member of the conference program committee

- Stéphane Girard organized an invited session "Regression extremes" at the *7th international conference ERCIM*, Pisa, Italy, december 2014.

9.1.1.4. member of the editorial board

- Florence Forbes is Associate Editor of the journal *Frontiers in ICT: Computer Image Analysis* since its creation in Sept. 2014. Computer Image Analysis is a new specialty section in the community-run openaccess journal Frontiers in ICT. This section is led by Specialty Chief Editors Drs Christian Barillot and Patrick Bouthemy.
- Stéphane Girard is Associate Editor of the *Statistics and Computing* journal since 2012. He is also member of the Advisory Board of the *Dependence Modelling* journal since decembre 2014.

9.1.1.5. reviewer

- In 2014, Florence Forbes has been a reviewer for the *NIPS* and *ICASSP* conferences and for the *Statistics and Computing* journal.
- In 2014, Stéphane Girard has been a reviewer for *Annals of Statistics, Journal of Statistical Software, Metrika, Lecture Notes in Statistics, RevStat, ESAIM Probability & Statistics, Journal de la Société Française de Statistique.*

9.1.2. Societies and Networks

- F. Forbes and J.-B. Durand are part of an INRA (French National Institute for Agricultural Research) Network (AIGM, http://carlit.toulouse.inra.fr/AIGM) on Algorithmic issues for inference in graphical models.
- F. Forbes and S. Girard were elected as members of the bureau of the "Analyse d'images, quantification, et statistique" group in the Société Française de Statistique (SFdS).
- F. Forbes and M-J. Martinez are members of the ERCIM working group on Mixture models.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence (IUT): Marie-José Martinez , *Statistics*, 192 ETD, L1 to L3 levels, université Grenoble 2, France.

Master: Jean-Baptiste Durand, *Statistics and probabilty*, 192 ETD, M1 and M2 levels, Ensimag Grenoble INP, France.

Licence (IUT) : Gildas Mazo, Mathematics and C language, 128h, L1 level, université Grenoble 1, France.

Master: Farida Enikeeva, *Statistics*, 96 ETD, M1 level, Ensimag Grenoble INP, France.

Master : Stéphane Girard, *Statistique Inférentielle Avancée*, 45 ETD, M1 level, Ensimag Grenoble-INP, France and *Introduction à la statistique des valeurs extrêmes*, 12 ETD, M2 level, université Grenoble 2, France.

Master : Florence Forbes, Mixture models and EM algorithm, 12h, M2 level, UFR IM2A, université Grenoble 1, France.

M.-J. Martinez is faculty members at Univ. Pierre Mendès France, Grenoble II.

J.-B. Durand is a faculty member at Ensimag, Grenoble INP.

F. Enikeeva was on a half-time ATER position at Ensimag, Grenoble INP.

9.2.2. Supervision

- PhD : Pierre Fernique, *"A statistical modeling framework for analyzing tree-indexed data"*, Montpellier 2 University. 10 Dec. 2014, Y. Guédon, J.-B. Durand.
- PhD : Gildas Mazo, *"Construction et estimation de copules en grande dimension"*, Universite Grenoble 1, 17 nov 2014, S. Girard, F. Forbes.

9.2.3. Juries

Stéphane Girard has been involved in the following PhD commitees:

- Blandine Fillon "Développement d'un outil statistique pour évaluer les charges maximales subies par l'isolation d'une cuve de méthanier au cours de sa période d'exploitation", Univ. Poitiers, December 2014.
- Tom Rohmer "Deux tests de détection de rupture dans la copule d'observations multivariées", Univ. Pau et des Pays de l'Adour, October 2014.
- Anthony Zullo "Functional analysis of high dimensional remote sensing images : application to the charaterization of semi-natural objects in landscape ecology", Univ. Toulouse, July 2014.

Florence Forbes has been involved in the PhD committees of:

- Haithem Boussaid, "Efficient Inference and learning in Graphical models for multi-organ shape segmentation", Ecole Centrale Paris, January 8, 2015 (President).
- Zacharie Irace, "Modelisation statistique et segmentation d'images TEP. Application à l'hétérogénéité et au suivi de tumeurs", INP Toulouse, Oct 8, 2014 (Reviewer).
- Vincent Brault, "Estimation et selection de modèle pour le modèle des blocs latents" , Paris-Sud University, Sept 9, 2014 (Reviewer).

Florence Forbes has been reviewer for the HDR committee of:

- Stéphane Chrétien, "Contribution à l'analyse et à l'amélioration de certaines méthodes pour l'inférence statistique par vraisemblance pénalisée", from Univeristy of Besancon, in Dec. 2014.

From Sept. 2009 to Sept. 2014, F. Forbes was head of the committee in charge of examining post-doctoral candidates at Inria Grenoble Rhône-Alpes ("Comité des Emplois Scientifiques").

Florence Forbes 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 was a member of the committee for research scientist candidate (CR) selection at Inria Lille and at Inria Grenoble in 2014.

10. 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^{o} 5, pp. 821-837 [*DOI :* 10.1109/TMI.2012.2225636], http://hal.inria.fr/inserm-00753873

[6] 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

[7] 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^{o} 3, pp. 824-837

[8] 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^{o} 6, pp. 971-984 [*DOI :* 10.1007/s11222-013-9414-4], https://hal.inria.fr/hal-00823451

[9] S. GIRARD. *A Hill type estimate of the Weibull tail-coefficient*, in "Communication in Statistics - Theory and Methods", 2004, vol. 33, n^{o} 2, pp. 205–234

[10] S. GIRARD, P. JACOB. *Extreme values and Haar series estimates of point process boundaries*, in "Scandinavian Journal of Statistics", 2003, vol. 30, n^{o} 2, pp. 369–384

Publications of the year

Articles in International Peer-Reviewed Journals

[11] B. BARROCA, P. BERNADARA, S. GIRARD, G. MAZO. *Considering hazard estimation uncertain in urban resilience strategies*, in "Natural Hazards and Earth System Sciences", 2015, vol. 15, pp. 25-34 [*DOI :* 10.5194/NHESS-15-25-2015], https://hal.archives-ouvertes.fr/hal-01100539

[12] C. BOUVEYRON, M. FAUVEL, S. GIRARD. *Kernel discriminant analysis and clustering with parsimonious Gaussian process models*, in "Statistics and Computing", 2014, 33 pages - arXiv:1204.4021, forthcoming, https://hal.archives-ouvertes.fr/hal-00687304

[13] M. CHAVENT, S. GIRARD, V. KUENTZ, B. LIQUET, T. M. N. NGUYEN, J. SARACCO. *A sliced inverse regression approach for data stream*, in "Computational Statistics", 2014, vol. 29, pp. 1129–1152 [*DOI :* 10.1007/s00180-014-0483-4], https://hal.inria.fr/hal-00688609

[14] M. CHINI, A. CHIANCONE, S. STRAMONDO. *Scale Object Selection (SOS) through a hierarchical segmentation by a multi-spectral per-pixel classification*, in "Pattern Recognition Letters", November 2014, vol. 49, pp. 214-223 [*DOI :* 10.1016/j.patrec.2014.07.012], https://hal.archives-ouvertes.fr/hal-01065938

[15] R. COUDRET, S. GIRARD, J. SARACCO. *A new sliced inverse regression method for multivariate response*, in "Computational Statistics and Data Analysis", September 2014, vol. 77, pp. 285-299 [*DOI :* 10.1016/j.csda.2014.03.006], https://hal.inria.fr/hal-00714981

[16] A. DAOUIA, S. GIRARD, A. GUILLOU. *A Gamma-moment approach to monotonic boundaries estimation: with applications in econometric and nuclear fields*, in "Journal of Econometrics", February 2014, vol. 178, n° 2, pp. 727-740 [*DOI :* 10.1016/j.jeconom.2013.10.013], https://hal.inria.fr/hal-00737732

[17] 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

[18] A. DELEFORGE, F. FORBES, R. HORAUD. *Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds*, in "International Journal of Neural Systems", February 2015, vol. 25, n° 1, 21 p. [*DOI :* 10.1142/S0129065714400036], https://hal.inria.fr/hal-00960796

[19] J.-B. DURAND, Y. GUÉDON. *Localizing the latent structure canonical uncertainty: entropy profiles for hidden Markov models*, in "Statistics and Computing", 2014, 19 p. [*DOI :* 10.1007/s11222-014-9494-9], https://hal.inria.fr/hal-01090836

[20] J. EL METHNI, L. GARDES, S. GIRARD. *Non-parametric estimation of extreme risk measures from conditional heavy-tailed distributions*, in "Scandinavian Journal of Statistics", 2014, vol. 41, n° 4, pp. 988–1012 [*DOI :* 10.1111/sjos.12078], https://hal.archives-ouvertes.fr/hal-00830647

[21] 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

[22] S. GIRARD, A. GUILLOU, G. STUPFLER. *Uniform strong consistency of a frontier estimator using kernel regression on high order moments*, in "ESAIM: Probability and Statistics", 2014, vol. 18, pp. 642–666 [*DOI :* 10.1051/ps/2013050], https://hal.archives-ouvertes.fr/hal-00764425

[23] M.-J. MARTINEZ, E. HOLIAN. *An alternative estimation approach for the heterogeneity linear mixed model*, in "Communications in Statistics - Simulation and Computation", 2014, vol. 43, n° 10, pp. 2628-2638 [*DOI :* 10.1080/03610918.2012.762389], https://hal.archives-ouvertes.fr/hal-00926620

[24] B. MENZE, A. JAKAB, S. BAUER, J. KALPATHY-CRAMER, K. FARAHANI, J. KIRBY, Y. BURREN, N. PORZ, J. SLOTBOOM, R. WIEST, L. LANCZI, E. GERSTNER, M.-A. WEBER, T. ARBEL, B. AVANTS, N. AYACHE, P. BUENDIA, L. COLLINS, N. CORDIER, J. CORSO, A. CRIMINISI, T. DAS, H. DELINGETTE, C. DEMIRALP, C. DURST, M. DOJAT, S. DOYLE, J. FESTA, F. FORBES, E. GEREMIA, B. GLOCKER, P. GOLLAND, X. GUO, A. HAMAMCI, K. IFTEKHARUDDIN, R. JENA, N. JOHN, E. KONUKOGLU, D.

LASHKARI, J. ANTONIO MARIZ, R. MEIER, S. PEREIRA, D. PRECUP, S. J. PRICE, T. RIKLIN-RAVIV, S. REZA, M. RYAN, L. SCHWARTZ, H.-C. SHIN, J. SHOTTON, C. SILVA, N. SOUSA, N. SUBBANNA, G. SZEKELY, T. TAYLOR, O. THOMAS, N. TUSTISON, G. UNAL, F. VASSEUR, M. WINTERMARK, D. HYE YE, L. ZHAO, B. ZHAO, D. ZIKIC, M. PRASTAWA, M. REYES, K. VAN LEEMPUT. *The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)*, in "IEEE Transactions on Medical Imaging", 2014, 33 p. [*DOI :* 10.1109/TMI.2014.2377694], https://hal.inria.fr/hal-00935640

[25] A. NAZIN, S. GIRARD. *L1-optimal linear programming estimatorfor periodic frontier functions with Holder continuous derivative*, in "Automation and Remote Control / Avtomatika i Telemekhanika", 2014, vol. 75, n^{o} 12, pp. 2152-2169, https://hal.archives-ouvertes.fr/hal-01066739

[26] T. VINCENT, S. BADILLO, L. RISSER, L. CHAARI, C. BAKHOUS, F. FORBES, P. CIUCIU. *Frontiers in Neuroinformatics Flexible multivariate hemodynamics fMRI data analyses and simulations with PyHRF*, in "Frontiers in Neuroscience", April 2014, vol. 8, n^{o} Article 67, 23 p. [*DOI :* 10.3389/FNINS.2014.00067], https://hal.inria.fr/hal-01084249

Invited Conferences

[27] F. FORBES, A. DELEFORGE, R. HORAUD. *High dimensional regression with Gaussian mixtures and partially latent response variables*, in "SuSTaIn Image Processing workshop "high-dimensional stochastic simulation and optimisation in image processing"", Bristol, United Kingdom, August 2014 [*DOI :* 10.1007/s11222-014-9461-5], https://hal.inria.fr/hal-01107604

[28] F. FORBES, A. FRAU-PASCUAL, T. VINCENT, J. SLOBODA, P. CIUCIU. *Physiologically informed Bayesian analysis of ASL fMRI data*, in "Statistical Challenges in Neuroscience workshop", Warwick, United Kingdom, September 2014, https://hal.inria.fr/hal-01107613

[29] F. FORBES, D. WRAITH. *Robust mixture modelling using skewed multivariate distributions with variable amounts of tailweight*, in "7th International Conference of the ERCIM WG on Computing and Statistics", Pise, Italy, October 2014, https://hal.inria.fr/hal-01107622

International Conferences with Proceedings

[30] M. ALBUGHDADI, L. CHAARI, F. FORBES, J.-Y. TOURNERET, P. CIUCIU. *Model Selection for Hemodynamic Brain Parcellation in fMRI*, in "EUSIPCO - 22nd European Signal Processing Conference", Lisbon, Portugal, IEEE, September 2014, pp. 31 - 35, https://hal.inria.fr/hal-01107475

[31] A. DELEFORGE, F. FORBES, R. HORAUD. *Hyper-spectral Image Analysis with Partially-Latent Regression*, in "22nd European Signal Processing Conference", Lisbon, Portugal, September 2014, https://hal.archives-ouvertes.fr/hal-01019360

[32] S. DOYLE, B. LEMASSON, F. VASSEUR, P. BOURDILLION, F. DUCRAY, J. HONNORAT, L. GUILLOTON, J. GUYOTAT, C. REMY, F. FORBES, F. COTTON, E. BARBIER, M. DOJAT. *Comparison of manual versus automatic delineation of low-grade gliomas based on MR brain scans*, in "Organization for Human Brain Mapping (OHBM) 2014 Annual meeting", Hambourg, Germany, June 2014, https://hal.inria.fr/hal-01107700

[33] J.-B. DURAND, Y. GUÉDON. *Quantifying and localizing state uncertainty in hidden Markov models using conditional entropy profiles*, in "COMPSTAT 2014 - 21st International Conference on Computational Statistics", Genève, Switzerland, M. GILLI, G. GONZÁLEZ-RODRÍGUEZ, A. NIETO-REYES (editors), Université

de Genève, August 2014, pp. 213-221, ISBN 978-2-8399-1347-8. Please check publisher, https://hal.inria.fr/hal-01058278

[34] J. EL METHNI, S. GIRARD, L. GARDES. *Kernel estimation of extreme risk measures for all domains of attraction*, in "COMPSTAT 2014 - 21st International Conference on Computational Statistics", Geneva, Switzerland, August 2014, CDROM, https://hal.archives-ouvertes.fr/hal-01062363

[35] M. FAUVEL, C. BOUVEYRON, S. GIRARD. *Parsimonious Gaussian Process Models for the Classification of Multivariate Remote Sensing Images*, in "ICASSP - IEEE International Conference on Acoustics, Speech, and Signal Processing", Florence, Italy, IEEE, May 2014, pp. 2913-2916 [*DOI :* 10.1109/ICASSP.2014.6854133], https://hal.archives-ouvertes.fr/hal-01062378

[36] P. FERNIQUE, J.-B. DURAND, Y. GUÉDON. *Estimation of Discrete Partially Directed Acyclic Graphical Models in Multitype Branching Processes*, in "COMPSTAT - 21st International Conference on Computational Statistics", Geneva, Switzerland, Proceedings of COMPSTAT 2014, The International Association for Statistical Computing (IASC), August 2014, https://hal.inria.fr/hal-01084524

[37] P. FERNIQUE, J.-B. DURAND, Y. GUÉDON. *Learning Discrete Partially Directed Acyclic Graphical Models in Multitype Branching Processes*, in "COMPSTAT 2014 - 21st International Conference on Computational Statistics", Genève, Switzerland, M. GILLI, G. GONZÁLEZ-RODRÍGUEZ, A. NIETO-REYES (editors), Université de Genève, August 2014, pp. 579-586, ISBN 978-2-8399-1347-8. Please check publisher, https://hal.inria.fr/hal-01058284

[38] A. FRAU-PASCUAL, T. VINCENT, F. FORBES, P. CIUCIU. *Hemodynamically informed parcellation of cerebral FMRI data*, in "ICASSP - IEEE International Conference on Acoustics, Speech, and Signal Processing", Florence, Italy, IEEE, May 2014, pp. 2079-2083 [*DOI :* 10.1109/ICASSP.2014.6853965], https://hal.inria.fr/hal-01100186

[39] A. FRAU-PASCUAL, T. VINCENT, J. SLOBODA, P. CIUCIU, F. FORBES. *Physiologically Informed Bayesian Analysis of ASL fMRI Data*, in "BAMBI 2014 - First International Workshop on Bayesian and grAphical Models for Biomedical Imaging", Boston, United States, M. J. CARDOSO, I. SIMPSON, T. ARBEL, D. PRECUP, A. RIBBENS (editors), Lecture Notes in Computer Science, Springer International Publishing, September 2014, vol. 8677, pp. 37 - 48 [*DOI :* 10.1007/978-3-319-12289-2_4], https://hal.inria.fr/hal-01100266

[40] I.-D. GEBRU, X. ALAMEDA-PINEDA, R. HORAUD, F. FORBES. *Audio-Visual Speaker Localization via Weighted Clustering*, in "IEEE Workshop on Machine Learning for Signal Processing", Reims, France, September 2014, 6 p. , https://hal.archives-ouvertes.fr/hal-01053732

[41] S. GIRARD, G. STUPFLER. *Extreme geometric quantiles*, in "7th International Conference of the ERCIM WG on Computing and Statistics", Pise, Italy, December 2014, https://hal.archives-ouvertes.fr/hal-01093048

[42] G. MAZO, S. GIRARD, F. FORBES. *A flexible, tractable class of copulas and its estimation*, in "COMPSTAT 2014 - 21st International Conference on Computational Statistics", Geneva, Switzerland, August 2014, https://hal.archives-ouvertes.fr/hal-01062481

[43] A. ROCHE, F. FORBES. *Partial volume estimation in brain MRI revisited*, in "MICCAI 2014 - 17th International Conference on Medical Image Computing and Computer Assisted Intervention", Boston, United

States, P. GOLLAND, N. HATA, C. BARILLOT, J. HORNEGGER, R. HOWE (editors), Springer, September 2014, vol. 8673, pp. 771-778 [*DOI :* 10.1007/978-3-319-10404-1_96], https://hal.inria.fr/hal-01107469

[44] G. STUPFLER, S. GIRARD. *On the asymptotic behaviour of extreme geometric quantiles*, in "Workshop on Extreme Value Theory, with an emphasis on spatial and temporal aspects", Besancon, France, November 2014, https://hal.archives-ouvertes.fr/hal-01086054

National Conferences with Proceedings

[45] A. ARNAUD, F. FORBES, N. COQUERY, E. BARBIER, B. LEMASSON. *Mélanges de lois de Student multivariées généralisées : application à la caractérisation de tumeurs par IRM multiparamétrique*, in "2ème congrès de la SFRMBM (Société Française de Résonance Magnétique en Biologie et Médecine)", Grenoble, France, March 2015, https://hal.inria.fr/hal-01107483

[46] A. CHIANCONE, S. GIRARD, J. CHANUSSOT. *Collaborative Sliced Inverse Regression*, in "Rencontres d'Astrostatistique", Grenoble, France, November 2014, https://hal.archives-ouvertes.fr/hal-01086931

[47] S. DOYLE, B. LEMASSON, F. VASSEUR, P. BOURDILLION, F. DUCRAY, J. HONNORAT, L. GUILLOTON, J. GUYOTAT, C. REMY, F. FORBES, F. COTTON, E. BARBIER, M. DOJAT. *Segmentation des tumeurs cérébrales de bas grade par une approche bayésienne : délinéation manuelle versus automatique*, in "2ème congrès de la SFRMBM (Société Française de Résonance Magnétique en Biologie et Médecine)", Grenoble, France, March 2015, https://hal.inria.fr/hal-01107520

[48] J.-B. DURAND, Y. GUÉDON. *Quantification de l'incertitude sur la structure latente dans des modèles de Markov cachés*, in "46èmes Journées de Statistique", Rennes, France, June 2014, https://hal.inria.fr/hal-01058317

[49] P. FERNIQUE, J.-B. DURAND, Y. GUÉDON. *Modèles graphiques paramétriques pour la modélisation des lois de génération dans des processus de branchement multitypes*, in "46èmes Journées de Statistique", Rennes, France, June 2014, https://hal.inria.fr/hal-01058313

[50] F. FORBES, A. DELEFORGE, R. HORAUD. *High dimensional regression with Gaussian mixtures and partially latent response variables: Application to hyper-spectral image analysis*, in "Rencontre d'Astrostatistique", Grenoble, France, November 2014, https://hal.inria.fr/hal-01107616

[51] S. SYLLA, S. GIRARD, A. K. DIONGUE, A. DIALLO, C. SOKHNA. *Classification supervisee par modele de melange: Application aux diagnostics par autopsie verbale*, in "46èmes Journées de Statistique organisées par la Société Française de Statistique", Rennes, France, June 2014, https://hal.archives-ouvertes.fr/hal-01090014

Scientific Books (or Scientific Book chapters)

[52] E. H. DEME, S. GIRARD, A. GUILLOU. *Reduced-bias estimator of the Conditional Tail Expectation of heavy-tailed distributions*, in "Mathematical Statistics and Limit Theorems", D. MASON (editor), Springer, 2014, https://hal.inria.fr/hal-00823260

[53] S. GIRARD, S. LOUHICHI. *On the strong consistency of the kernel estimator of extreme conditional quantiles*, in "Recent advances in statistical methodology and its application", E. O. SAID (editor), Springer, 2014, https://hal.inria.fr/hal-01058390

[54] M.-J. MARTINEZ, J. HINDE. *Random effects ordinal time models for grouped toxicological data from a biological control assay*, in "Statistical Modelling in Biostatistics and Bioinformatics: Selected papers", G. MACKENZIE, D. PENG (editors), Contributions to Statistics, Springer, May 2014, pp. 45-58 [*DOI :* 10.1007/978-3-319-04579-5_5], https://hal.archives-ouvertes.fr/hal-00943962

Other Publications

[55] C. AMBLARD, S. GIRARD, L. MENNETEAU. *Bivariate copulas defined from matrices*, 2014, https://hal.archives-ouvertes.fr/hal-00875303

[56] C. BAZZOLI, F. LETUÉ, M.-J. MARTINEZ. *Modelling finger force produced from different tasks using linear mixed models with lme R function*, June 2014, https://hal.archives-ouvertes.fr/hal-00998910

[57] L. GARDES, S. GIRARD. *Nonparametric estimation of the conditional tail copula*, March 2014, https://hal.archives-ouvertes.fr/hal-00964514

[58] L. GARDES, S. GIRARD. *On the estimation of the functional Weibull tail-coefficient*, 2014, https://hal.archives-ouvertes.fr/hal-01063569

[59] S. GIRARD. *An introduction to SIR: A statistical method for dimension reduction in multivariate regression*, 2014, https://hal.archives-ouvertes.fr/hal-01058721

[60] S. GIRARD, S. LOUHICHI. *On the strong consistency of the kernel estimator of extreme conditional quantiles*, March 2014, https://hal.archives-ouvertes.fr/hal-00956351

[61] S. GIRARD, J. SARACCO. *An introduction to dimension reduction in nonparametric kernel regression*, 2014, https://hal.archives-ouvertes.fr/hal-00977512

[62] S. GIRARD, G. STUPFLER. *Asymptotic behaviour of extreme geometric quantiles and their estimation under moment conditions*, September 2014, https://hal.inria.fr/hal-01060985

[63] S. GIRARD, G. STUPFLER. *Intriguing properties of extreme geometric quantiles*, February 2014, https://hal.inria.fr/hal-00865767

[64] G. MAZO, S. GIRARD, F. FORBES. *A class of multivariate copulas based on products of bivariate copulas*, July 2014, https://hal.archives-ouvertes.fr/hal-00910775

[65] G. MAZO, S. GIRARD, F. FORBES. *A flexible and tractable class of one-factor copulas*, April 2014, https://hal.archives-ouvertes.fr/hal-00979147

[66] G. MAZO, S. GIRARD, F. FORBES. *Weighted least-squares inference based on dependence coefficients for multivariate copulas*, April 2014, https://hal.archives-ouvertes.fr/hal-00979151

References in notes

[67] 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

[68] J.-B. DURAND, P. GONÇALVÈS, Y. GUÉDON. *Computational Methods for Hidden Markov Tree Models – An Application to Wavelet Trees*, in "IEEE Transactions on Signal Processing", September 2004, vol. 52, n^{o} 9, pp. 2551–2560

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

[70] P. FERNIQUE. *A statistical modeling framework for analyzing tree-indexed data*, Université de Montpellier 2, December 2014, http://tel.archives-ouvertes.fr/tel-01095420

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

[72] 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

[73] D. KOLLER, N. FRIEDMAN. *Probabilistic graphical models: principles and techniques*, MIT press, 2009

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

[75] R. NELSEN. *An introduction to copulas*, Lecture Notes in Statistics, Springer-VerlagNew-York, 1999, vol. 139

[76] 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

[77] T. VINCENT, J. WARNKING, M. VILLIEN, A. KRAINIK, P. CIUCIU, F. FORBES. *Bayesian Joint Detection-Estimation of cerebral vasoreactivity from ASL fMRI data*, in "MICCAI 2013 - 16th International Conference on Medical Image Computing and Computer Assisted Intervention", Nagoya, Japan, K. MORI, I. SAKUMA, Y. SATO, C. BARILLOT, N. NAVAB (editors), Lecture Notes in Computer Science, Springer, June 2013, vol. 8150, pp. 616-623 [*DOI :* 10.1007/978-3-642-40763-5_76], http://hal.inria.fr/hal-00854437