IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2012

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, 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: January 01, 2008 .

1. Members

Research Scientists
Florence Forbes [Team Leader, DR,Inria, HdR]
Stéphane Girard [CR, Inria, HdR]

Faculty Members
Jean-Baptiste Durand [INPG, Grenoble]
Marie-José Martinez [UPMF, Grenoble]

Engineers
Senan James Doyle [Inria]
Ludovic Leau-Mercier [Inria]
Darren Wraith [Inria]

PhD Students
Jonathan El-Methni [Inria, from October 2010, co-advised by L. Gardes and S. Girard]
El-Hadji Deme [Université Gaston Berger, Sénégal, March-May, 2012]
Seydou-Nourou Sylla [Université Gaston Berger, Sénégal, October-December, 2012]
Christine Bakhous [Inria, from November 2010, co-advised by F. Forbes and M. Dojat (GIN)]
Gildas Mazo [Inria, from October 2011, co-advised by F. Forbes and S. Girard]

Post-Doctoral Fellows
Kai Qin [Inria, until June 2012]
Angelika Studeny [Inria, since Sept. 2012]
Lotfi Chaari [Inria, until August 2012]
Huu Giao Nguyen [Inria]
Farida Enikeeva [UJF, Grenoble]
Thomas Vincent [Inria]

Administrative Assistant
Imma Presseguer [Inria]

Others
Marc Guillotin [Intern, Inria, April-June. 2012]
Minwoo Lee [Intern, Inria, June-August 2012]

2. Overall Objectives

2.1. Introduction

The MISTIS team aims to develop statistical methods for dealing with complex problems or data. Our applications consist mainly of image processing and spatial data problems with some applications in biology and medicine. Our approach is based on the statement that complexity can be handled by working up from simple local assumptions in a coherent way, defining a structured model, and that is the key to modelling, computation, inference and interpretation. 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 statistical methodology (finite mixture analysis, hidden Markov models, random effect models, etc). 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. The focus is on functions describing curves or surfaces or more generally manifolds rather than real valued parameters. This can be interesting in image processing for instance where it can be difficult to introduce parametric models that are general enough (e.g. for contours).

2.2. Highlights of the Year

Our paper [33] entitled *An Improved CUDA-Based Implementation of Differential Evolution on GPU* was nominated and finalist for the best paper award in the Digital Entertainment Technologies and Arts / Parallel Evolutionary Systems session of the Genetic and Evolutionary Computation Conference 2012 (Gecco 2012).

3. Scientific Foundations

3.1. Mixture models

Participants: Angelika Studeny, Thomas Vincent, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Jean-Baptiste Durand, Florence Forbes, Stéphane Girard, Marie-José Martinez, Darren Wraith.

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 are 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, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Jean-Baptiste Durand, Florence Forbes, Darren Wraith.

graphical models, Markov properties, conditional independence, hidden Markov trees, clustering, statistical learning, missing data, mixture of distributions, EM algorithm, stochastic algorithms, selection and combination of models, statistical pattern recognition, image analysis, hidden Markov field, Bayesian inference

Graphical modelling provides a diagrammatic representation of the logical 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 the mean field principle 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: El-Hadji Deme, Jonathan El-Methni, Ludovic Leau-Mercier, Stéphane Girard, Gildas Mazo, Kai Qin, Huu Giao Nguyen, Farida Enikeeva, Seydou-Nourou Sylla.

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 [59]. 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 [68] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [58], 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).

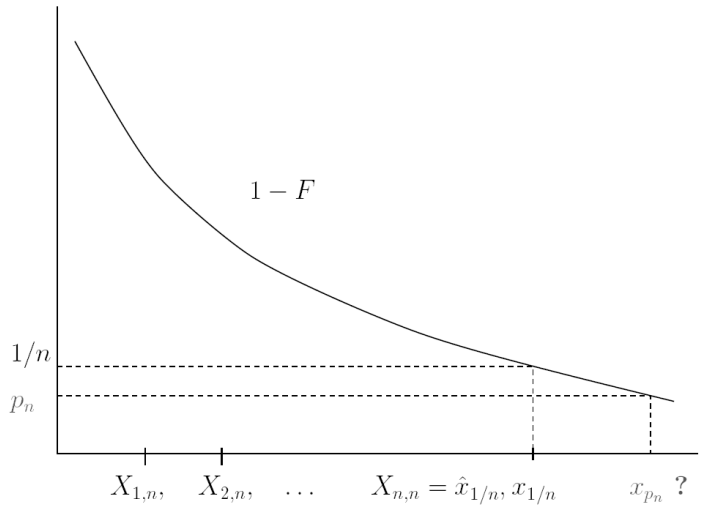

Figure 1. The curve represents the survival function $x \to P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\widehat{x}_{1/n} = X_{n,n}$. As illustrated in the figure, to estimate p_n-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.

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 [62]. 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 [53]. Such models can then be used in a Mixtures or Markov framework for classification purposes. Another approach consists in combining dimension reduction, regularization techniques, and regression techniques to improve the Sliced Inverse Regression method [68].

4. Application Domains

4.1. Image Analysis

Participants: Christine Bakhous, Lotfi Chaari, Senan James Doyle, Thomas Vincent, Florence Forbes, Ludovic Leau-Mercier, Huu Giao Nguyen, Stéphane Girard, Kai Qin, Darren Wraith.

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 LNAO from 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. In the context of the ANR MDCO project Vahine, we work on hyperspectral multi-angle images. In the context of the "pole de competivite" project I-VP, we work of images of PC Boards.

4.2. Biology, Environment and Medicine

Participants: Thomas Vincent, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Florence Forbes, Stéphane Girard, Jonathan El-Methni, Gildas Mazo, Angelika Studeny.

A second domain of applications concerns biology and medecine. 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 population genetics and neurosiences are also considered. Finally, in the context of the ANR VMC project Medup, we study the uncertainties on the forecasting and climate projection for Mediterranean high-impact weather events.

4.3. Reliability

Participants: Jean-Baptiste Durand, Stéphane Girard.

Reliability and industrial lifetime analysis are applications developed through collaborations with the EDF research department and the LCFR laboratory (Laboratoire de Conduite et Fiabilité des Réacteurs) of CEA / Cadarache. We also consider failure detection in print infrastructure [16] through collaboration with Xerox, Meylan.

5. Software

5.1. The ECMPR software

Participant: Florence Forbes.

Joint work with: Radu Horaud and Manuel Iguel.

The ECMPR (Expectation Conditional Maximization for Point Registration) package implements [57] [65]. It registers two (2D or 3D) point clouds using an algorithm based on maximum likelihood with hidden variables. The method can register both rigid and articulated shapes. It estimates both the rigid or the kinematic transformation between the two shapes as well as the parameters (covariances) associated with the underlying Gaussian mixture model. It has been registered in APP in 2010 under the GPL license.

5.2. The LOCUS and P-LOCUS software

Participants: Florence Forbes, Senan James Doyle.

Joint work with: Michel Dojat.

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 and its recent extension P-LOCUS automatically perform this segmentation for healthy and pathological 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. Its extension (P-LOCUS) for lesion detection is realized by S. Doyle with financial support from Gravit for possible industrial transfer.

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.

5.3. The POPEYE software

Participant: Florence Forbes.

Joint work with: Vasil Khalidov, Radu Horaud, Miles Hansard, Ramya Narasimha, Elise Arnaud.

POPEYE contains software modules and libraries jointly developed by three partners within the POP STREP project: Inria, University of Sheffield, and University of Coimbra. It includes kinematic and dynamic control of the robot head, stereo calibration, camera-microphone calibration, auditory and image processing, stereo matching, binaural localization, audio-visual speaker localization. Currently, this software package is not distributed outside POP.

5.4. The HDDA and HDDC toolboxes

Participant: Stéphane Girard.

Joint work with: Charles Bouveyron (Université Paris 1). The High-Dimensional Discriminant Analysis (HDDA) and the High-Dimensional Data Clustering (HDDC) toolboxes contain respectively efficient supervised and unsupervised classifiers for high-dimensional data. These classifiers are based on Gaussian models adapted for high-dimensional data [53]. The HDDA and HDDC toolboxes are available for Matlab and are included into the software MixMod [52]. Recently, a R package has been developed and integrated in The Comprehensive R Archive Network (CRAN). It can be downloaded at the following URL: http://cran.r-project.org/web/packages/HDclassif/.

5.5. The Extremes freeware

Participant: Stéphane Girard.

Joint work with: Diebolt, J. (CNRS), Laurent Gardes (Univ Strasbourg) and Garrido, M. (INRA Clermont-Ferrand-Theix).

The EXTREMES software is a toolbox dedicated to the modelling of extremal events offering extreme quantile estimation procedures and model selection methods. This software results from a collaboration with EDF R&D. It is also a consequence of the PhD thesis work of Myriam Garrido [55]. The software is written in C++ with a Matlab graphical interface. It is now available both on Windows and Linux environments. It can be downloaded at the following URL: http://extremes.gforge.inria.fr/.

5.6. The SpaCEM^3 program

Participants: Senan James Doyle, Florence Forbes.

SpaCEM^3 (Spatial Clustering with EM and Markov Models) is a software that provides a wide range of supervised or unsupervised clustering algorithms. The main originality of the proposed algorithms is that clustered objects do not need to be assumed independent and can be associated with very high-dimensional measurements. Typical examples include image segmentation where the objects are the pixels on a regular grid and depend on neighbouring pixels on this grid. More generally, the software provides algorithms to cluster multimodal data with an underlying dependence structure accounting for some spatial localisation or some kind of interaction that can be encoded in a graph.

This software, developed by present and past members of the team, is the result of several research developments on the subject. The current version 2.09 of the software is CeCILLB licensed.
Main features. The approach is based on the EM algorithm for clustering and on Markov Random Fields (MRF) to account for dependencies. In addition to standard clustering tools based on independent Gaussian mixture models, SpaCEM^3 features include:

- The unsupervised clustering of dependent objects. Their dependencies are encoded via a graph not necessarily regular and data sets are modelled via Markov random fields and mixture models (eg. MRF and Hidden MRF). Available Markov models include extensions of the Potts model with the possibility to define more general interaction models.
- The supervised clustering of dependent objects when standard Hidden MRF (HMRF) assumptions do not hold (ie. in the case of non-correlated and non-unimodal noise models). The learning and test steps are based on recently introduced Triplet Markov models.
- Selection model criteria (BIC, ICL and their mean-field approximations) that select the "best" HMRF according to the data.
- The possibility of producing simulated data from:
  - general pairwise MRF with singleton and pair potentials (typically Potts models and extensions)
  - standard HMRF, ie. with independent noise model
  - general Triplet Markov models with interaction up to order 2
- A specific setting to account for high-dimensional observations.
- An integrated framework to deal with missing observations, under Missing At Random (MAR) hypothesis, with prior imputation (KNN, mean, etc), online imputation (as a step in the algorithm), or without imputation.

The software is available at http://spacem3.gforge.inria.fr. A user manual in English is available on the web site above together with example data sets. The INRA Toulouse unit is more recently participating to this project for promotion among the bioinformatics community [75].

5.7. The FASTRUCT software

Participant: Florence Forbes.

Joint work with: Francois, O. (TimB, TIMC) and Chen, C. (former Post-doctoral fellow in Mistis).

The FASTRUCT program is dedicated to the modelling and inference of population structure from genetic data. Bayesian model-based clustering programs have gained increased popularity in studies of population structure since the publication of the software STRUCTURE [70]. These programs are generally acknowledged as performing well, but their running-time may be prohibitive. FASTRUCT is a non-Bayesian implementation of the classical model with no-admixture uncorrelated allele frequencies. This new program relies on the Expectation-Maximization principle, and produces assignment rivaling other model-based clustering programs. In addition, it can be several-fold faster than Bayesian implementations. The software consists of a command-line engine, which is suitable for batch-analysis of data, and a MS Windows graphical interface, which is convenient for exploring data.

It is written for Windows OS and contains a detailed user's guide. It is available at http://mistis.inrialpes.fr/realisations.html.

The functionalities are further described in the related publication:

- Molecular Ecology Notes 2006 [56].

5.8. The TESS software

Participant: Florence Forbes.

Joint work with: Francois, O. (TimB, TIMC) and Chen, C. (former post-doctoral fellow in Mistis).

TESS is a computer program that implements a Bayesian clustering algorithm for spatial population genetics. Is it particularly useful for seeking genetic barriers or genetic discontinuities in continuous populations. The method is based on a hierarchical mixture model where the prior distribution on cluster labels is defined as a Hidden Markov Random Field [60]. Given individual geographical locations, the program seeks population structure from multilocus genotypes without assuming predefined populations. TESS takes input data files in a format compatible to existing non-spatial Bayesian algorithms (e.g. STRUCTURE). It returns graphical displays of cluster membership probabilities and geographical cluster assignments through its Graphical User Interface.

The functionalities and the comparison with three other Bayesian Clustering programs are specified in the following publication:

- Molecular Ecology Notes 2007

6. New Results

6.1. Mixture models

6.1.1. Taking into account the curse of dimensionality

Participant: Stéphane Girard.

Joint work with: Bouveyron, C. (Université Paris 1), Fauvel, M. (ENSAT Toulouse)

In the PhD work of Charles Bouveyron (co-advised by Cordelia Schmid from the Inria LEAR team) [53], 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]. Also, the description of the R package is published in [11]. Our recent work consists in adding a kernel in the previous methods to deal with nonlinear data classification [27], [45].

6.1.2. Robust mixture modelling using skewed multivariate distributions with variable amounts of tailweight

Participants: Florence Forbes, Darren Wraith.

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). In such cases, the multivariate student t distribution is motivated as a heavy-tailed alternative to the multivariate Gaussian distribution. The additional flexibility of the multivariate t comes from introducing an additional degree of freedom parameter (dof) which can be viewed as a robust tuning parameter.

A useful representation of the t-distribution is as a so-called *infinite mixture of scaled Gaussians* or *Gaussian scale mixture*,

$$p(y;\mu,\Sigma,\theta) = \int_0^\infty \mathcal{N}_M(y;\mu,\Sigma/w)\ f_W(w;\theta)\ \mathrm{d}w \tag{5}$$

where $\mathcal{N}_M(\ .\ ;\mu,\Sigma/w)$ denotes the M-dimensional Gaussian distribution with mean μ and covariance Σ/w and f_W is the probability distribution of a univariate positive variable W referred to as the weight variable. When f_W is a Gamma distribution $\mathcal{G}(\nu/2,\nu/2)$ where ν denotes the degrees of freedom, we recover the multivariate t distribution. The weight variable W in this case effectively acts to govern the tail behaviour of the distributional form from light tails ($\nu \to \infty$) to heavy tails ($\nu \to 0$) depending on the value of ν.

For many applications, the distribution of the data may also be highly asymmetric in addition to being heavy tailed (or affected by outliers). A natural extension to the Gaussian scale mixture case is to consider *location **and** scale Gaussian mixtures* of the form,

$$p(y;\mu,\Sigma,\theta) = \int_0^\infty \mathcal{N}_M(y;\mu+w\beta\Sigma,w\Sigma)\ f_W(w;\theta)\ \mathrm{d}w, \tag{6}$$

where β is an additional M-dimensional vector parameter for skewness and the determinant of Σ equals 1 for parameter identifiability. When f_W is a Generalized Inverse Gaussian distribution ($GIG(y;\lambda,\delta,\gamma)$, we recover the family of Generalized Hyperbolic (GH) distributions. Depending on the parameter choice for the GIG, special cases of the GH family, include: the multivariate GH distribution with hyperbolic margins ($\lambda = 1$); the normal inverse Gaussian distribution ($\lambda = -1/2$); the multivariate hyperbolic ($\lambda = \frac{M+1}{2}$) distribution; the hyperboloid distribution ($\lambda = 0$); the hyperbolic skew-t distribution ($\lambda = -\nu, \gamma = 0$); and the normal gamma distribution ($\lambda > 0, \mu = 0, \delta = 0$)) amongst others. For applied problems, the most popular of these forms appears to be the Normal Inverse Gaussian (NIG) distribution, with extensive use in financial applications. Another distributional form allowing for skewness and heavy or light tails includes different forms of the multivariate skew-t. Most of these distributional forms are also able to be represented as *location and scale Gaussian mixtures*.

Although the above approaches provide for great flexibility in modelling data of highly asymmetric and heavy tailed form the above approaches assume f_W to be a univariate distribution and hence each dimension is governed by the same amount of tailweight. There have been various approaches to address this issue in the statistics literature for both symmetric and asymmetric distributional forms. In his work, [66] proposes a dependent bivariate t-distribution with marginals of different degrees of freedom but the tractability of the extension to the multivariate case is unclear. Additional proposals are reviewed in chapters 4 and 5 of [67] but these formulations tend to be appreciably more complicated, often already in the expression of the probability density function. Increasingly, there has been much research on copula approaches to account for flexible distributional forms but the choice as to which one to use in this case and the applicability to (even) moderate dimensions is also not clear. In general the papers take various approaches whose relationships have been characterized in the bivariate case by [73]. However, most of the existing approaches suffer either from the non-existence of a closed-form pdf or from a difficult generalization to more than two dimensions.

In this work, we show that the location and scale mixture representation can be further explored and propose a framework that is considerably simpler than those previously proposed with distributions exhibiting interesting properties. Using the normal inverse Gaussian distribution (NIG) as an example, we extend the standard *location and scale mixture of Gaussian representation* to allow for the tail behaviour to be set or estimated differently in each dimension of the variable space. The key elements of the approach are the introduction of multidimensional weights and a decomposition of the matrix Σ in (6) which facilitates the separate estimation and also allows for arbitrary correlation between dimensions. We outline an approach for maximum likelihood estimation of the parameters via the EM algorithm and explore the performance of the approach on several simulated and real data sets in the context of clustering.

6.1.3. Robust clustering for high dimensional data

Participants: Florence Forbes, Darren Wraith, Minwoo Lee.

For a clustering problem, a parametric mixture model is one of the popular approaches. Most of all, Gaussian mixture models are widely used in various fields of study such as data mining, pattern recognition, machine learning, and statistical analysis. The modeling and computational flexibility of the Gaussian mixture model makes it possible to model a rich class of density, and provides a simple mathematical form of cluster models.

Despite the success of Gaussian mixtures, the parameter estimations can be severely affected by outliers. By adding an additional degrees of freedom (dof) parameter, a robustness tuning parameter, the robust improvement in clustering has been achieved. Although adopting t distribution loses the closed-form solution, it is still tractable by representing t distribution as Gaussian scale mixture (GSM), which consists of a Gaussian random vector that is weighted by a hidden scaling variable. Recent work that uses the multivariate t distribution has showed the improved robustness.

Along with robustness from t distribution, for the practical use, efficient handling of a high dimensional data is critical. High dimensional data often make most of clustering methods perform poorly. To overcome the curse of dimensionality, Bouveyron et al. [54] proposed the model-based high dimensional data clustering (HDDC). HDDC searches the intrinsic dimension of each class with the BIC criterion or the scree-test of Cattell; this allows them to limit the number of parameters by taking into account only the specific subspace that each class is located. The parameterization makes HDDC not only computationally efficient but robust with respect to the ill-conditioning or the singularity of empirical covariance matrix.

This work proposes an approach that combines robust clustering with the HDDC. The use of the mixture of multivariate t distribution on the basis of HDDC develops robust high dimensional clustering methods that can capture various kinds of density models. Further, extending the mixture model with multiple t distributions for each dimension, we propose more flexible model that can be applicable to various data. We suggest a model-based approach for this method.

6.1.4. Partially Supervised Mapping: A Unified Model for Regression and Dimensionality Reduction

Participant: Florence Forbes.

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

We cast dimensionality reduction and regression in a unified latent variable model. We propose a two-step strategy consisting of characterizing a non-linear *reversed* output-to-input regression with a generative piecewise-linear model, followed by Bayes inversion to obtain an output density given an input. We describe and analyze the most general case of this model, namely when only some components of the output variables are observed while the other components are latent. We provide two EM inference procedures and their initialization. Using simulated and real data, we show that the proposed method outperforms several existing ones.

6.1.5. Variational EM for Binaural Sound-Source Separation and Localization

Participant: Florence Forbes.

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

We addressed the problem of sound-source separation and localization in real-world conditions with two microphones. Both tasks are solved within a unified formulation using supervised mapping. While the parameters of the direct mapping are learned during a training stage that uses sources emitting white noise (calibration), the inverse mapping is estimated using a variational EM formulation. The proposed algorithm can deal with natural sound sources such as speech which are known to yield sparse spectrograms, and is able to locate multiple sources both in azimuth and in elevation. Extensive experiments with real data show that the method outperform state-of-the-art both in separation and localization.

6.2. Statistical models for Neuroscience

6.2.1. Variational approach for the joint estimation-detection of Brain activity from functional MRI data

Participants: Florence Forbes, Lotfi Chaari, Thomas Vincent.

Joint work with: Michel Dojat (Grenoble Institute of Neuroscience) and Philippe Ciuciu from Neurospin, CEA in Saclay.

In standard within-subject analyses of event-related fMRI data, two steps are usually performed separately: detection of brain activity and estimation of the hemodynamic response. Because these two steps are inherently linked, we adopt the so-called region-based Joint Detection-Estimation (JDE) framework that addresses this joint issue using a multivariate inference for detection and estimation. JDE is built by making use of a regional bilinear generative model of the BOLD response and constraining the parameter estimation by physiological priors using temporal and spatial information in a Markovian model. In contrast to previous works that use Markov Chain Monte Carlo (MCMC) techniques to sample the resulting intractable posterior distribution, we recast the JDE into a missing data framework and derive a Variational Expectation-Maximization (VEM) algorithm for its inference. A variational approximation is used to approximate the Markovian model in the unsupervised spatially adaptive JDE inference, which allows automatic fine-tuning of spatial regularization parameters. It provides a new algorithm that exhibits interesting properties terms of estimation error and computational cost compared to the previously used MCMC-based approach. Experiments on artificial and real data show that VEM-JDE is robust to model mis-specification and provides computational gain while maintaining good performance in terms of activation detection and hemodynamic shape recovery. Main corresponding paper [13]

6.2.2. Hemodynamic-informed parcellation of fMRI data in a Joint Detection Estimation framework

Participants: Florence Forbes, Lotfi Chaari, Thomas Vincent.

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

Identifying brain hemodynamics in event-related functional MRI (fMRI) data is a crucial issue to disentangle the vascular response from the neuronal activity in the BOLD signal. This question is usually addressed by estimating the so-called Hemodynamic Response Function (HRF). Voxelwise or region-/parcelwise inference schemes have been proposed to achieve this goal but so far all known contributions commit to pre-specified spatial supports for the hemodynamic territories by defining these supports either as individual voxels or a priori fixed brain parcels. In this paper, we introduce a Joint Parcellation-Detection-Estimation (JPDE) procedure that incorporates an adaptive parcel identification step based upon local hemodynamic properties. Efficient inference of both evoked activity, HRF shapes and *supports* is then achieved using variational approximations. Validation on synthetic and real fMRI data demonstrate the JPDE performance over standard detection estimation schemes and suggest it as a new brain exploration tool. Corresponding papers [29], [28].

6.2.3. Variational variable selection to assess experimental condition relevance in event-related fMRI

Participants: Florence Forbes, Christine Bakhous, Lotfi Chaari, Thomas Vincent, Farida Enikeeva.

Joint work with: Michel Dojat (Grenoble Institute of Neuroscience) and Philippe Ciuciu from Neurospin, CEA in Saclay.

Brain functional exploration investigates the nature of neural processing following cognitive or sensory stimulation. This goal is not fully accounted for in most functional Magnetic Resonance Imaging (fMRI) analysis which usually assumes that all delivered stimuli possibly generate a BOLD response everywhere in the brain although activation is likely to be induced by only some of them in specific brain regions. Generally, criteria are not available to select the relevant conditions or stimulus types (e.g. visual, auditory, etc.) prior to activation detection and the inclusion of irrelevant events may degrade the results, particularly when the Hemodynamic Response Function (HRF) is jointly estimated. To face this issue, we propose an efficient variational procedure that automatically selects the conditions according to the brain activity they elicit. It follows an improved activation detection and local HRF estimation that we illustrate on synthetic and real fMRI data. This approach is an alternative to our previous approach based on Monte-Carlo Markov Chain (MCMC) inference [25]. Corresponding paper [26].

6.2.4. Bayesian BOLD and perfusion source separation and deconvolution from functional ASL imaging

Participants: Florence Forbes, Thomas Vincent.

In the context of ARC AINSI project, joint work with: Philippe Ciuciu from Neurospin, CEA in Saclay.

In many neuroscience applications, the Arterial Spin Labeling (ASL) fMRI modality arises as a preferable choice to the standard BOLD modality due to its ability to provide a quantitative measure of the Cerebral Blood Flow (CBF). Such a quantification is central but generally performed without consideration of a specific modeling of the perfusion component in the signal often handled via standard GLM approaches using the BOLD canonical response function as regressor. In this work, we propose a novel Bayesian hierarchical model of the ASL signal which allows activation detection and both the extraction of a perfusion and a hemodynamic component. Validation on synthetic and real data sets from event-related ASL show the ability of our model to address the source separation and double deconvolution problems inherent to ASL data analysis.

6.2.5. Extraction of physiological components in functional ASL data

Participants: Florence Forbes, Thomas Vincent, Lotfi Chaari, Marc Guillotin.

In the context of ARC AINSI project, joint work with: Jan Warnking (Grenoble Institute of Neuroscience) and Philippe Ciuciu from Neurospin, CEA in Saclay.

The internship of Marc Guillotin has been supported by Le pole Cognition de Grenoble.

The goal of this work was to investigate Independent component analysis techniques to identify the part of the ASL signal due to physiological sources such as respiratory and cardiac components. Once identified those physiological components should be removed to produce an uncontaminated ASL signal. This preliminary work showed that the physiological effects were affecting all signal components and were therefore not easy to extract without removing some of the useful signal. More experiments should be made on real data from the GIN.

6.2.6. Comparison of processing workflows for ASL data analysis

Participant: Thomas Vincent.

In the context of ARC AINSI project, joint work with: Michel Dojat (Grenoble Institute of Neuroscience), Philippe Ciuciu from Neurospin, CEA in Saclay, Remi Dubujet, Elise Bannier, Isabelle Courouge, Christian Barillot, Camille Maudet from EPI Visages in Rennes.

We assessed and compared the performance of different ASL processing pipelines in order to promote one using specific indexes (Contrast to noise ratio, partial volume effect, et). We proposed to assess the impact of the pipelines based on the quality of the final corrected ASL images using a common set of subjects for all workflows. We leaned on the expertise of the Visages and GIN teams on ASL, and first started from existing attempts made in the teams. At the moment, there is a striking lack of such guidelines. The recent toolbox ASLtbx proposes a number of procedures that are based on very standard tools (e.g. SPM) and do not make use of more efficient approaches from more recent literature. Similarly, in the BIRN project, processing pipelines are mentioned but none are currently available.

6.3. Markov models

6.3.1. Spatial risk mapping for rare disease with hidden Markov fields and variational EM

Participants: Florence Forbes, Senan James Doyle.

Joint work with: Lamiae Azizi, David Abrial and Myriam Garrido from INRA Clermont-Ferrand-Theix.

Current risk mapping models for pooled data focus on the estimated risk for each geographical unit. A risk classification, *i.e.* grouping of geographical units with similar risk, is then necessary to easily draw interpretable maps, with clearly delimited zones in which protection measures can be applied. As an illustration, we focus on the Bovine Spongiform Encephalopathy (BSE) disease that threatened the bovine production in Europe and generated drastic cow culling. This example features typical animal disease risk analysis issues with very low risk values, small numbers of observed cases and population sizes that increase the difficulty of an automatic classification. We propose to handle this task in a spatial clustering framework using a non standard discrete hidden Markov model prior designed to favor a smooth risk variation. The model parameters are estimated using an EM algorithm and a mean field approximation for which we develop a new initialization strategy appropriate for spatial Poisson mixtures. Using both simulated and our BSE data, we show that our strategy performs well in dealing with low population sizes and accurately determines high risk regions, both in terms of localization and risk level estimation.

Main corresponding paper [14].

6.3.2. Spatial modelling of biodiversity from high-througput DNA sequence data

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 EPI Bamboo

Biodiversity has been acknowledged as a vital ressource for ecosystem health and stability, faced with an unprecedented global decline. In order to be effective, conservation actions need to be based on reliable and fast analysis. Recent advances in DNA sequencing methods now enable DNA-based identification of multiple species from only few, even potentially degraded environmental samples (metabarcoding.org, [74]). This offers a new way of biodiversity assessment and is of particular interest where large-scale individual-based diversity assessment is difficult, for example in tropical environments. Due to their comparatively low demand in cost and effort, these methods are characterized by their high throughput; they are expected to produce vast amounts of data as they gain in popularity over the coming years. The specific properties of these data (e.g. bias from sequencing errors, notion of species) and their high dimensionality provides new statistical and computational challenges for biodiversity assessment. This project aims at extending existing summary statistics to be used with data from metabarcoding surveys and, where this is not adequate, to develop new methodology. A special focus is on the spatial mapping of biodiversity and the co-occurrence of species. In a first instance, we investigate spatial clustering algorithms based on Markov random fields (software SpaCEM3, http://spacem3.gforge.inria.fr/) to identify regions of high species occurrence as well as structured additive regression models and their implementation to estimate cross-correlations between species occurrences in space [61], [72], [71]. At present, results have been derived in form of species occurrence maps, which take into account pairwise cross-correlation, and interaction graphs.

6.3.3. 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.

The quantity and quality of yields in fruit trees is closely related to processes of growth and branching, which determine ultimately the regularity of flowering and the position of flowers. Flowering and fruiting patterns are explained by statistical dependence between the nature of a parent shoot (*e.g.* flowering or not) and the quantity and natures of its children shoots – with potential effect of covariates. Thus, better characterization of patterns and dependencies is expected to lead to strategies to control the demographic properties of the shoots (through varietal selection or crop management policies), and thus to bring substantial improvements in the quantity and quality of yields.

Since the connections between shoots can be represented by mathematical trees, statistical models based on multitype branching processes and Markov trees appear as a natural tool to model the dependencies of interest. Formally, the properties of a vertex are summed up using the notion of vertex state. In such models, the numbers of children in each state given the parent state are modeled through discrete multivariate distributions. Model selection procedures are necessary to specify parsimonious distributions. We developed an approach based on probabilistic graphical models 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 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 were chosen by selection model procedures among different parametric families.

This work was carried out in the context of Pierre Fernique's first year of PhD (Montpellier 2 University and CIRAD). It was applied to model dependencies between short or long, vegetative or flowering shoots in apple trees. The results highlighted contrasted patterns related to the parent shoot state, with interpretation in terms of alternation of flowering (see paragraph 6.3.4). It was also applied to the analysis of the connections between cyclic growth and flowering of mango trees. This work will be continued during Pierre Fernique's PhD thesis, with extensions to other fruit tree species and other parametric discrete multivariate families of distributions, including covariates and mixed effects.

6.3.4. Statistical characterization of the alternation of flowering in fruit tree species

Participant: Jean-Baptiste Durand.

Joint work with: Jean Peyhardi and Yann Guédon (Mixed Research Unit DAP, Virtual Plants team), Baptiste Guitton, Yan Holtz and Evelyne Costes (DAP, AFEF team), Catherine Trottier (Montpellier University)

The aim of this work was to characterize genetic determinisms of the alternation of flowering in apple tree progenies. Data were collected at two scales: at whole tree scale (with annual time step) and a local scale (annual shoot or AS, which is the portions of stem that were grown during the same year). Two replications of each genotype were available.

Indices were proposed to characterize alternation at tree scale. The difficulty is related to early detection of alternating genotypes, in a context where alternation is often concealed by a substantial increase of the number of flowers over consecutive years. To separate correctly the increase of the number of flowers due to aging of young trees from alternation in flowering, our model relied on a parametric hypothesis for the trend (fixed slopes specific to genotype and random slopes specific to replications), which translated into mixed effect modelling. Then, different indices of alternation were computed on the residuals. Clusters of individuals with contrasted patterns of bearing habits were identified.

To model alternation of flowering at AS scale, a second-order Markov tree model was built. Its transition probabilities were modelled as generalized linear mixed models, to incorporate the effects of genotypes, year and memory of flowering for the Markovian part, with interactions between these components.

Asynchronism of flowering at AS scale was assessed using an entropy-based criterion. The entropy allowed for a characterisation of the roles of local alternation and asynchronism in regularity of flowering at tree scale.

Moreover, our models highlighted significant correlations between indices of alternation at AS and individual scales.

This work was extended by the Master 2 internship of Yan Holtz, supervised by Evelyne Costes and Jean-Baptiste Durand. New progenies were considered, and a methodology based on a lighter measurement protocol was developed and assessed. It consisted in assessing the accuracy of approximating the indices computed from measurements at tree scale by the same indices computed as AS scale. The approximations were shown sufficiently accurate to provide an operational strategy for apple tree selection.

As a perspective of this work, patterns in the production of children ASs (numbers of flowering and vegetative children) depending on the type of the parent AS must be analyzed using branching processes and different types of Markov trees, in the context of Pierre Fernique's PhD Thesis (see paragraph 6.3.3).

6.4. Semi and non-parametric methods

6.4.1. Post-Reflow Automated Optical Inspection of Lead Defects

Participants: Florence Forbes, Kai Qin, Huu Giao Nguyen, Darren Wraith, Ludovic Leau-mercier.

This is joint work with VI-Technology in the context of the IVP project.

Quality and throughput in printed circuit board (PCB) assembly lines constitute a continuous challenge, especially when placing smaller components on boards that are becoming increasingly dense. Automated optical inspection (AOI) technology allows PCB assembly lines to keep operating at a high throughput while visually inspecting production quality in term of paste deposits, mounted components and solder joints in an automatic and non-contact manner. In the AOI, high definition cameras precisely move in both X- and Y-direction to scan the device under test lit by special lighting techniques, e.g. light-emitting diode (LED) lighting. The captured images are then analyzed using specific inspection algorithms to identify defects. The AOI systems can be placed at several stages during the manufacturing process, such as bare board inspection, solder paste inspection, pre-reflow inspection and post-reflow inspection, which usually need some time to be programmed via offline learning of verified boards and expert expertise before online inspection starts. Vi TECHNOLOGY (VIT) offers a wide range of AOI solutions to increase productivity throughout electronics manufacturing lines while enhancing the quality of products. Post-reflow AOI is implemented after the reflow procedure in PCB assembly lines to enable inspection of the major post-reflow defects. This work focus on certain types of post-reflow defects occurring on leaded components, i.e. lifted lead, no solder, excess of solder, contamination on lead, insufficient solder, bad wedding and dry joint. We aim at developing efficient post-reflow lead defect detection approaches by synergizing image analysis, pattern recognition, machine learning, and statistics techniques to improve performance of VIT commercial post-reflow AOI solutions from two aspects: 1) Reducing both detection escape rate and false detection rate; 2) Minimizing programming efforts. The exact nature of the work is confidential.

6.4.2. An Improved CUDA-Based Implementation of Differential Evolution on GPU

Participants: Kai Qin, Florence Forbes.

Modern GPUs enable widely affordable personal computers to carry out massively parallel computation tasks. NVIDIA's CUDA technology provides a wieldy parallel computing platform. Many state-of-the-art algorithms arising from different fields have been redesigned based on CUDA to achieve computational speedup. Differential evolution (DE), as a very promising evolutionary algorithm, is highly suitable for parallelization owing to its data parallel algorithmic structure. However, most existing CUDA based DE implementations suffer from excessive low-throughput memory access and less efficient device utilization. This work presents an improved CUDA-based DE to optimize memory and device utilization: several logically-related kernels are combined into one composite kernel to reduce global memory access; kernel execution configuration parameters are automatically determined to maximize device occupancy; streams are employed to enable concurrent kernel execution to maximize device utilization. Experimental results on several numerical problems demonstrate superior computational time efficiency of the proposed method over two recent CUDA-based DE and the sequential DE across varying problem dimensions and algorithmic population sizes.

This work was nominated for the best paper award (finalist) in the Digital Entertainment Technologies and Arts / Parallel Evolutionary Systems session of the Genetic and Evolutionary Computation Conference 2012 (GECCO12) conference [33].

6.4.3. Augmented cumulative distribution networks for multivariate extreme value modelling

Participants: Stéphane Girard, Gildas Mazo, Florence Forbes.

Max-stable distribution functions are theoretically grounded models for modelling multivariate extreme values. However they suffer from some striking limitations when applied to real data analysis due to the intractability of the likelihood when the number of variables becomes high. Cumulative Distribution Networks (CDN's) have been introduced recently in the machine learning community and allow the construction of max-stable distribution functions for which the density can be computed. Unfortunately, we show in this work that the dependence structure expected in the data may not be accurately reflected by max-stable CDN's. To face this limitation, we therefore propose to augment max-stable CDN's with the more standard Gumbel max-stable distribution function in order to enrich the dependence structure [32].

6.4.4. Modelling extremal events

Participants: Stéphane Girard, Jonathan El-Methni, El-Hadji Deme.

Joint work with: Guillou, A. and Gardes, L. (Univ. Strasbourg).

We introduced a new model of tail distributions depending on two parameters $\tau \in [0,1]$ and $\theta > 0$. This model includes very different distribution tail behaviors from Fréchet and Gumbel maximum domains of attraction. In the particular cases of Pareto type tails ($\tau = 1$) or Weibull tails ($\tau = 0$), our estimators coincide with classical ones proposed in the literature, thus permitting us to retrieve their asymptotic normality in an unified way. The first year of the PhD work of Jonathan El-methni has been dedicated to the definition of an estimator of the parameter τ. This permits the construction of new estimators of extreme quantiles. The results are published in [17]. Our future work will consist in proposing a test procedure in order to discriminate between Pareto and Weibull tails.
We are also working on the estimation of the second order parameter ρ (see paragraph 3.3.1). We proposed a new family of estimators encompassing the existing ones (see for instance [64], [63]). This work is in collaboration with El-Hadji Deme, a PhD student from the Université de Saint-Louis (Sénégal). El-Hadji Deme obtained a one-year mobility grant to work within the Mistis team on extreme-value statistics. The results are submitted for publication [49]. We also proposed reduced-bias estimators of the Proportional Hazard Premium for heavy-tailed distributions. The results are submitted for publication [50].

6.4.5. Conditional extremal events

Participants: Stéphane Girard, Gildas Mazo, Jonathan El-methni.

Joint work with: L. Gardes, Amblard, C. (TimB in TIMC laboratory, Univ. Grenoble I) and Daouia, A. (Univ. Toulouse I and Univ. Catholique de Louvain)

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 [59] with extreme-value methods in order to obtain efficient estimators of the conditional tail index and conditional extreme quantiles. When the covariate is functional and random (random design) and the tail of the distribution is heavy, we focus on kernel methods [18]. We extension to all kind of tails in investigated in [15].

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.

More future work will include the study of multivariate and spatial extreme values. With this aim, a research on some particular copulas [1] has been initiated with Cécile Amblard, since they are the key tool for building multivariate distributions [69]. The PhD theses of Jonathan El-methni and Gildas Mazo should address this issue too.

6.4.6. *Level sets estimation*

Participant: Stéphane Girard.

Joint work with: Guillou, A. and Gardes, L. (Univ. Strasbourg), Stupfler, G. (Univ. Strasbourg) and Daouia, A. (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].
In collaboration with A. Daouia, we investigate the application of such methods in econometrics [42], [48]: 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 the PhD thesis of Gilles Stupfler (co-directed by Armelle Guillou and Stéphane Girard), 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 [19], [20], [21].

6.4.7. *Quantifying uncertainties on extreme rainfall estimations*

Participant: Stéphane Girard.

Joint work with: Carreau, J. (Hydrosciences Montpellier), Gardes, L. (univ. Strasbourg) and Molinié, G. from Laboratoire d'Etude des Transferts en Hydrologie et Environnement (LTHE), France.

Extreme rainfalls are generally associated with two different precipitation regimes. Extreme cumulated rainfall over 24 hours results from stratiform clouds on which the relief forcing is of primary importance. Extreme rainfall rates are defined as rainfall rates with low probability of occurrence, typically with higher mean return-levels than the maximum observed level. For example Figure 2 presents the return levels for the Cévennes-Vivarais region that can be obtained. It is then of primary importance to study the sensitivity of the extreme rainfall estimation to the estimation method considered.

The obtained results are published in [12].

6.4.8. *Retrieval of Mars surface physical properties from OMEGA hyperspectral images.*

Participant: Stéphane Girard.

Joint work with: Douté, S. from Laboratoire de Planétologie de Grenoble, France and Saracco, J (University 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) [47]. We have also defined an adaptive version of the method which is able to deal with block-wise evolving data streams [46].

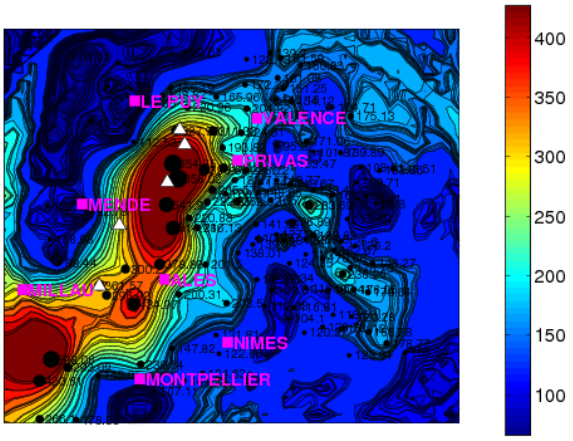

Figure 2. Map of the mean return-levels (in mm) for a period of 10 years.

6.4.9. Statistical modelling development for low power processor.

Participant: Stéphane Girard.

Joint work with: A. Lombardot and S. Joshi (ST Crolles).

With scaling down technologies to the nanometer regime, the static power dissipation in semiconductor devices is becoming more and more important. Techniques to accurately estimate System On Chip static power dissipation are becoming essential. Traditionally, designers use a standard corner based approach to optimize and check their devices. However, this approach can drastically underestimate or over-estimate process variations impact and leads to important errors.

The need for an effective modeling of process variation for static power analysis has led to the introduction of Statistical static power analysis. Some publication state that it is possible to save up to 50% static power using statistical approach. However, most of the statistical approaches are based on Monte Carlo analysis, and such methods are not suited to large devices. It is thus necessary to develop solutions for large devices integrated in an industrial design flow. Our objective to model the total consumption of the circuit from the probability distribution of consumption of each individual gate. Our preliminary results are published in [23].

7. Partnerships and Cooperations

7.1. Regional Initiatives

MISTIS participates in the weekly statistical seminar of Grenoble. F. Forbes is one of the organizers 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.

7.2. National Initiatives

7.2.1. Competitivity Clusters

MISTIS is a partner in a three-year (2010-12) MINALOGIC project (I-VP for Intuitive Vision Programming) supported by the French Government. The project is led by VI Technology (http://www.vitechnology.com),

a world leader in Automated Optical Inspection (AOI) of a broad range of electronic components. The other partners involved are the CMM (Centre de Morphologie Mathematiques) in Fontainebleau, and Pige Electronique in Bourg-Les-Valence. The overall goal is to exploit statistical and image processing techniques more intensively to improve defect detection capability and programming time based on existing AOI principles so as to eventually reach a reliable defect detection with virtually zero programming skills and efforts.

7.2.2. ARC Inria

Florence Forbes is coordinating the 2-year Inria ARC project AINSI (http://thalie.ujf-grenoble.fr/ainsi). AINSI stands for "Modeles statistiques pour l'Assimilation d'Informations de Neuroimagerie fonctionnelle et de perfuSIon cerebrale". The goal is to propose an innovative statistically well-based solution to the joint determination of neural activity and brain vascularization by combining BOLD constrast images obtained in functional MRI and quantitative parametric images (Arterial Spin Labelling: ASL). The partners involved are Visages team from Inria in Rennes and Parietal in Saclay, the INSERM Unit U594 (Grenoble Institute of Neuroscience) and the LNAO laboratory from CEA NeuroSpin.

7.3. European Initiatives

7.3.1. FP7 Projects

7.3.1.1. HUMAVIPS

Title: Humanoids with audiovisual skills in populated spaces

Type: COOPERATION (ICT)

Defi: Cognitive Systems and Robotics

Instrument: Specific Targeted Research Project (STREP)

Duration: February 2010 - January 2013

Coordinator: Inria (France)

Others partners: CTU Prague (Czech Republic), University of Bielefeld (Germany), IDIAP (Switzerland), Aldebaran Robotics (France)

See also: http://humavips.inrialpes.fr

Abstract: Humanoids expected to collaborate with people should be able to interact with them in the most natural way. This involves significant perceptual, communication, and motor processes, operating in a coordinated fashion. Consider a social gathering scenario where a humanoid is expected to possess certain social skills. It should be able to explore a populated space, to localize people and to determine their status, to decide to join one or two persons, to synthetize appropriate behavior, and to engage in dialog with them. Humans appear to solve these tasks routinely by integrating the often complementary information provided by multi sensory data processing, from low-level 3D object positioning to high-level gesture recognition and dialog handling. Understanding the world from unrestricted s

7.4. International Research Visitors

7.4.1. Internships

MINWOO JAKE LEE (from Jun 2012 until Aug 2012)

Subject: Clustering or classification of high dimensional data in the presence of outliers

Institution: Colorado State University (United States)

El Hadji DEME (from Mar 2012 until May 2012)

Subject: Bias reduction in extreme-value statistics

Institution: Université Gaston Berger (Senegal)

Seydou-Nourou Sylla (from October 2012 to December 2012)

Subject: Classification for medical data

Institution: Université Gaston Berger (Senegal)

8. Dissemination

8.1. Scientific Animation

Since September 2009, F. Forbes is head of the committee in charge of examining post-doctoral candidates at Inria Grenoble Rhône-Alpes ("Comité des Emplois Scientifiques").

Since September 2009, F. Forbes is also a member of the Inria national committee, "Comité d'animation scientifique", in charge of analyzing and motivating innovative activities in Applied Mathematics.

Florence Forbes is a member of an INRA committee (CSS MBIA) in charge of evaluating INRA researchers once a year.

F. Forbes is part of an INRA (French National Institute for Agricultural Research) Network (MSTGA) on spatial statistics.

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).

S. Girard is associate editor of the international journal "Statistics and Computing".

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Stéphane Girard

Master : Statistique inférentielle avancée, 45h, M1, Ensimag (Grenoble INP), France.

Florence Forbes

Master : Mixture models and EM algorithm, 12h, M2, UFR IM2A, Université Grenoble I, 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 is on a half-time ATER position at Ensimag, Grenoble INP.

C. Bakhous and J. El Methni are both moniteur at University Joseph Fourier.

PhD & HdR

PhD in progress : Jonathan El Methni, Modèles en statistique des valeurs extrêmes, since October, 2010, Stéphane Girard

PhD in progress : Christine Bakhous, Problèmes de sélection de modèles en IRM fonctionnelle, since November, 2010, Florence Forbes and Michel Dojat

PhD in progress : Gildas Mazo, Modèles spatiaux en statistique des valeurs extrèmes, since October, 2011, Florence Forbes and Stéphane Girard

PhD in progress : El Hadji Deme, Réduction du biais en statistique des valeurs extrèmes, since October, 2009, Stéphane Girard

PhD in progress : Seydou-Nourou Sylla, Modélisation statistique pour l'analyse de causes de décès décrites par autopsie verbale en milieu rural africain, since October, 2012, Stéphane Girard

8.2.2. Juries

Stéphane Girard was a member of the Strasbourg university committee in charge of examining applications for assistant professor in 2012.

Florence Forbes was also a member of an INRA committee in charge of examining applications for junior researcher positions in 2012 at dept MBIA of INRA.

F. Forbes was involved in the PhD committees of

- El Ghali Lazrak from Inria Nancy and INRA Aster Mirecourt, Université de Lorraine in October 2012 (reviewer).
- Alexandre Janon from Inria team MOISE and LJK Grenoble. November 2012 (Examinateur).
- Mahdi Bagher from Inria team Maverick and LJK Grenoble. November 2012 (Examinateur).

F. Forbes was also involved in the HDR committee of Michael Blum, CR CNRS at TimC in Grenoble. December 2012 (Examinateur).

9. 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, p. 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), p. 1055–1067.

[3] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional data clustering*, in "Computational Statistics and Data Analysis", 2007, vol. 52, p. 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] G. CELEUX, S. CHRÉTIEN, F. FORBES, A. MKHADRI. *A Component-wise EM Algorithm for Mixtures*, in "Journal of Computational and Graphical Statistics", 2001, vol. 10, p. 699–712.

[6] G. CELEUX, F. FORBES, N. PEYRARD. *EM procedures using mean field-like approximations for Markov model-based image segmentation*, in "Pattern Recognition", 2003, vol. 36, n^{o} 1, p. 131-144.

[7] F. FORBES, G. FORT. *Combining Monte Carlo and Mean field like methods for inference in hidden Markov Random Fields*, in "IEEE trans. PAMI", 2007, vol. 16, n^{o} 3, p. 824-837.

[8] F. FORBES, N. PEYRARD. *Hidden Markov Random Field Model Selection Criteria based on Mean Field-like Approximations*, in "IEEE trans. PAMI", August 2003, vol. 25(9), p. 1089–1101.

[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, p. 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, p. 369–384.

Publications of the year

Articles in International Peer-Reviewed Journals

[11] L. BERGÉ, C. BOUVEYRON, S. GIRARD. *HDclassif: an R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data*, in "Journal of Statistical Software", 2012, vol. 46, n^o 6, p. 1–29, http://hal.inria.fr/hal-00541203.

[12] J. CARREAU, D. CERESETTI, E. URSU, S. ANQUETIN, J. CREUTIN, L. GARDES, S. GIRARD, G. MOLINIÉ. *Evaluation of classical spatial-analysis schemes of extreme rainfall*, in "Natural Hazards and Earth System Sciences", 2012, vol. 12, p. 3229–3240.

[13] 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", October 2012 [*DOI :* 10.1109/TMI.2012.2225636], http://hal.inria.fr/inserm-00753873.

[14] M. CHARRAS-GARRIDO, L. AZIZI, F. FORBES, S. DOYLE, N. PEYRARD, D. ABRIAL. *On the difficulty to clearly identify and delineate disease risk hot spots*, in "International Journal of Applied Earth Observation and Geoinformation", May 2012, available on line.

[15] A. DAOUIA, L. GARDES, S. GIRARD. *On kernel smoothing for extremal quantile regression*, in "Bernoulli", 2013, to appear.

[16] J. DURAND, S. GIRARD, V. CIRIZA, L. DONINI. *Optimization of power consumption and user impact based on point process modeling of the request sequence*, in "Journal of the Royal Statistical Society series C", 2013, to appear.

[17] J. EL METHNI, L. GARDES, S. GIRARD, A. GUILLOU. *Estimation of extreme quantiles from heavy and light tailed distributions*, in "Journal of Statistical Planning and Inference", 2012, vol. 142, n^o 10, p. 2735-2747, http://hal.inria.fr/hal-00627964.

[18] L. GARDES, S. GIRARD. *Functional kernel estimators of large conditional quantiles*, in "Electronic Journal of Statistics", 2012, vol. 6, p. 1715-1744, http://hal.inria.fr/hal-00608192.

[19] S. GIRARD, A. GUILLOU, G. STUPFLER. *Estimating an endpoint with high order moments in the Weibull domain of attraction*, in "Statistics and Probability Letters", December 2012, vol. 82, p. 2136-2144, http://hal.inria.fr/hal-00648435.

[20] S. GIRARD, A. GUILLOU, G. STUPFLER. *Estimating an endpoint with high order moments*, in "Test", 2012, vol. 21, n^o 4, p. 697–729, http://hal.inria.fr/inria-00596979.

[21] S. GIRARD, A. GUILLOU, G. STUPFLER. *Frontier estimation with kernel regression on high order moments*, in "Journal of Multivariate Analysis", 2013, vol. 116, p. 172–189.

[22] C. HATT, F. MANKESSI, J.-B. DURAND, F. BOUDON, F. MONTES, M. LARTAUD, J.-L. VERDEIL, O. MONTEEUIS. *Characteristics of Acacia mangium shoot apical meristems in natural and in vitro conditions in relation to heteroblasty*, in "Trees - Structure and Function", 2012, vol. 26, n^o 3, p. 1031-1044, PDF version of the authors can be published in January 2013 [*DOI :* 10.1007/s00468-012-0680-0], http://hal.inria.fr/hal-00699815.

[23] S. JOSHI, A. LOMBARDOT, P. FLATRESSE, C. D'AGOSTINO, A. JUGE, E. BEIGNE, S. GIRARD. *Statistical estimation of dominant physical parameters for leakage variability in 32nanometer CMOS under supply voltage variations*, in "Journal of Low Power Electronics", 2012, vol. 8, p. 113–124.

International Conferences with Proceedings

[24] C. AMBLARD, S. GIRARD, L. MENNETEAU. *Algebraic properties of copulas defined from matrices*, in "Workshop on Copulae in Mathematical and Quantitative Finance", Cracovie, Pologne, juillet 2012.

[25] C. BAKHOUS, F. FORBES, T. VINCENT, L. CHAARI, M. DOJAT, P. CIUCIU. *Adaptive experimental condition selection in event-related fMRI*, in "ISBI 2012 - IEEE International Symposium on Biomedical Imaging", Barcelone, Spain, IEEE, May 2012, p. 1755-1758 [*DOI :* 10.1109/ISBI.2012.6235920], http://hal.inria.fr/cea-00710489.

[26] C. BAKHOUS, F. FORBES, T. VINCENT, M. DOJAT, P. CIUCIU. *Variational variable selection to assess experimental condition relevance in event-related fMRI*, in "ISBI 2013 - IEEE International Symposium on Biomedical Imaging", San Francisco, USA, IEEE, May 2013.

[27] C. BOUVEYRON, M. FAUVEL, S. GIRARD. *Kernel discriminant analysis and clustering with parsimonious Gaussian process models*, in "ICML workshop on Object, functional and structured data : towards next generation kernel-based methods", Edinburgh, United Kingdom, 2012, http://hal.inria.fr/hal-00707056.

[28] L. CHAARI, F. FORBES, T. VINCENT, P. CIUCIU. *Hemodynamic-informed parcellation of fMRI data in a Joint Detection Estimation framework*, in "International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)", Nice, France, October 1-5 2012.

[29] L. CHAARI, F. FORBES, T. VINCENT, P. CIUCIU. *Robust voxel-wise Joint Detection Estimation of brain activity in fMRI*, in "IEEE International Conference on Image Processing (ICIP)", Orlando, USA, September 30-October 3 2012.

[30] A. DAOUIA, L. GARDES, S. GIRARD. *On kernel smoothing for extremal quantile regression*, in "5th International Conference of the ERCIM WG on computing and statistics", Oviedo, Spain, décembre 2012.

[31] S. JOSHI, A. LOMBARDOT, M. BELLEVILLE, E. BEIGNE, S. GIRARD. *Statistical leakage estimation in 32nm CMOS considering cells correlations*, in "11th IEEE conference on Faible Tension Faible Consommation", Paris, juin 2012.

[32] G. MAZO, F. FORBES, S. GIRARD. *Augmented cumulative distribution networks for multivariate extreme value modelling*, in "5th International Conference of the ERCIM WG on Computing and Statistics", Oviedo, Spain, December 2012.

[33] K. QIN, F. RAIMONDO, F. FORBES, Y. S. ONG. *An Improved CUDA-Based Implementation of Differential Evolution on GPU*, in "Genetic and Evolutionary Computation Conference 2012 (Gecco 2012)", July 12-16 2012.

National Conferences with Proceeding

[34] C. BAKHOUS, F. FORBES, T. VINCENT, L. CHAARI, M. DOJAT, P. CIUCIU. *Sélection de variable dans un cadre bayésien de traitement de données d'IRM fonctionnelle*, in "Journées de Statistique de la Société Française de Statistique (SFdS)", Brussels, Belgium, May 21-25 2012.

[35] C. BOUVEYRON, M. FAUVEL, S. GIRARD. *Processus gaussiens parcimonieux pour la classification générative de données hétérogènes*, in "44èmes Journées de Statistique de la Société Française de Statistique", Bruxelles, Belgium, 2012, http://hal.inria.fr/hal-00707059.

[36] L. CHAARI, F. FORBES, P. CIUCIU, T. VINCENT. *Parcel-free Joint Detection-Estimation in fMRI*, in "Journées de Statistique de la Société Française de Statistique (SFdS)", Brussels, Belgium, May 21-25 2012.

[37] J. EL-METHNI, L. GARDES, S. GIRARD. *Estimation de l'espérance conditionnelle des pertes extrêmes dans le cas de lois à queues lourdes en présence d'une covariable*, in "44èmes Journées de Statistique organisées par la Société Française de Statistique", Bruxelles, Belgique, mai 2012.

[38] S. GIRARD, A. GUILLOU, G. STUPFLER. *Estimation de point terminal dans le domaine d'attraction de Weibull par une méthode des moments d'ordre élevé*, in "44èmes Journées de Statistique organisées par la Société Française de Statistique", Bruxelles, Belgique, mai 2012.

[39] J. SARACCO, M. CHAVENT, B. LIQUET, V. KUENTZ, T. NGUYEN, S. GIRARD. *Régression inverse par tranches sur flux de données*, in "44èmes Journées de Statistique organisées par la Société Française de Statistique", Bruxelles, Belgique, mai 2012.

Conferences without Proceedings

[40] L. BERGÉ, C. BOUVEYRON, S. GIRARD. *HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data*, in "1ères Rencontres R", Bordeaux, France, July 2012, http://hal.inria.fr/hal-00717506.

[41] L. GARDES, S. GIRARD. *Functional kernel estimators of conditional extreme quantiles*, in "7èmes Journées de Statistique Fonctionnelle et Opératorielle", Montpellier, juin 2012.

Scientific Books (or Scientific Book chapters)

[42] A. DAOUIA, L. GARDES, S. GIRARD. *Nadaraya's estimates for large quantiles and free disposal support curves*, in "Exploring research frontiers in contemporary statistics and econometrics", I. V. KEILEGOM, P. WILSON (editors), Springer, 2012, p. 1-22, http://hal.inria.fr/hal-00528670.

Research Reports

[43] J.-B. DURAND, Y. GUÉDON. *Localizing the Latent Structure Canonical Uncertainty: Entropy Profiles for Hidden Markov Models*, Inria, February 2012, n^{o} RR-7896, 43, Submitted to Journal of Machine Learning Research, http://hal.inria.fr/hal-00675223.

[44] V. KHALIDOV, F. FORBES, R. HORAUD. *Calibration of A Binocular-Binaural Sensor Using a Moving Audio-Visual Target*, Inria, January 2012, n^{o} RR-7865, 27, http://hal.inria.fr/hal-00662306.

Other Publications

[45] C. BOUVEYRON, M. FAUVEL, S. GIRARD. *Kernel discriminant analysis and clustering with parsimonious Gaussian process models*, http://hal.inria.fr/hal-00687304.

[46] M. CHAVENT, S. GIRARD, V. KUENTZ, B. LIQUET, T. M. N. NGUYEN, J. SARACCO. *A sliced inverse regression approach for data stream*, http://hal.inria.fr/hal-00688609.

[47] R. COUDRET, S. GIRARD, J. SARACCO. *A new sliced inverse regression method for multivariate response regression*, http://hal.inria.fr/hal-00714981.

[48] A. DAOUIA, S. GIRARD, A. GUILLOU. *A Gamma-moment approach to monotonic boundaries estimation: with applications in econometric and nuclear fields*, http://hal.inria.fr/hal-00737732.

[49] E. DEME, L. GARDES, S. GIRARD. *On the estimation of the second order parameter for heavy-tailed distributions*, 2012, http://hal.inria.fr/hal-00634573/fr/.

[50] E. H. DEME, S. GIRARD, A. GUILLOU. *Reduced-bias estimator of the Proportional Hazard Premium for heavy-tailed distributions*, 2012, http://hal.inria.fr/hal-00763978.

[51] S. GIRARD, A. GUILLOU, G. STUPFLER. *Uniform strong consistency of a frontier estimator using kernel regression on high order moments*, http://hal.inria.fr/hal-00764425.

References in notes

[52] C. BIERNACKI, G. CELEUX, G. GOVAERT, F. LANGROGNET. *Model-Based Cluster and Discriminant Analysis with the MIXMOD Software*, in "Computational Statistics and Data Analysis", 2006, vol. 51, n^{o} 2, p. 587–600.

[53] 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.

[54] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High-dimensional data clustering*, in "Computational Statistics & Data Analysis", 2007, vol. 52, n^{o} 1, p. 502–519.

[55] M. CHARRAS-GARRIDO. *Modélisation des événements rares et estimation des quantiles extrêmes, méthodes de sélection de modèles pour les queues de distribution*, Université Grenoble 1, juin 2002, http://mistis.inrialpes.fr/people/girard/Fichiers/theseGarrido.pdf.

[56] C. CHEN, F. FORBES, O. FRANCOIS. *FASTRUCT: Model-based clustering made faster*, in "Molecular Ecology Notes", 2006, vol. 6, p. 980–983.

[57] G. DEWAELE, F. DEVERNAY, R. HORAUD, F. FORBES. *The alignment between 3D-data and articulated shapes with bending surfaces*, in "European Conf. Computer Vision, Lecture notes in Computer Science", 2006, n^{o} 3, p. 578-591.

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

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

[60] O. FRANCOIS, S. ANCELET, G. GUILLOT. *Bayesian clustering using Hidden Markov Random Fields in spatial genetics*, in "Genetics", 2006, p. 805–816.

[61] A. E. GELFAND, A. M. SCHMIDT, S. BANERJEE, C. SIRMANS. *Nonstationary multivariate process modeling through spatially varying coregionalization*, in "Test", 2004, vol. 13, p. 263-312, http://dx.doi.org/10.1007/BF02595775.

[62] 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.

[63] Y. GOEGEBEUR, J. BEIRLANT, T. DE WET. *Kernel estimators for the second order parameter in extreme value statistics*, in "Journal of Statistical Planning and Inference", 2010, vol. 140, n^{o} 9, p. 2632–2652.

[64] M. GOMES, L. DE HAAN, L. PENG. *Semi-parametric Estimation of the Second Order Parameter in Statistics of Extremes*, in "Extremes", 2002, vol. 5, n^{o} 4, p. 387–414.

[65] R. HORAUD, F. FORBES, M. YGUEL, G. DEWAELE, J. ZHANG. *Rigid and Articulated Point Registration with Expectation Conditional Maximization*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", March 2011, vol. 33, n^{o} 3, p. 587–602 [*DOI :* 10.1109/TPAMI.2010.94], http://hal.inria.fr/inria-00590265/en.

[66] M. JONES. *A dependent bivariate t distribution with marginals on different degrees of freedom*, in "Statistics and Probability Letters", 2002, vol. 56, n^{o} 2, p. 163-170.

[67] S. KOTZ, S. NADARAJAH. *Multivariate t Distributions and their Applications*, Cambridge, 2004.

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

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

[70] J. PRITCHARD, M. STEPHENS, P. DONNELLY. *Inference of Population Structure Using Multilocus Genotype Data*, in "Genetics", 2000, vol. 155, p. 945–959.

[71] H. RUE, S. MARTINO, N. CHOPIN. *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion)*, in "Journal of the Royal Statistical Society B", 2009, vol. 71, p. 319–392.

[72] A. M. SCHMIDT, M. A. RODRIGUEZ. *Modelling multivariate counts varying continuously in space*, in "Bayesian Statistics", J. M. BERNARDO, M. J. BAYARRI, J. O. BERGER, A. P. DAWID, D. HECKERMAN, A. F. M. SMITH, M. WEST (editors), Oxford University Press, 2010.

[73] W. T. SHAW, K. T. A. LEE. *Bivariate Student distributions with variable marginal degrees of freedom and independence*, in "Journal of Multivariate Analysis", 2008, vol. 99, n^{o} 6, p. 1276-1287.

[74] P. TABERLET, E. COISSAC, M. HAJIBABAEI, L. RIESEBERG. *Environmental DNA*, 2011, vol. 21, Molecular Ecology, special issue.

[75] M. VIGNES, J. BLANCHET, D. LEROUX, F. FORBES. *SpaCEM3, a software for biological module detection when data is incomplete, high dimensional and dependent*, in "Bioinformatics", 2011, vol. 27, n^{o} 6, p. 881-882.