INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team MISTIS

Modelling and Inference of Complex and Structured Stochastic Systems

Rhône-Alpes

THEME COG

Table of contents

1. Team

Team leader
Florence Forbes [Research scientist, Inria]

Research scientists
Stéphane Girard [Research scientist, Inria, HdR]
Laurent Gardes [Faculty member, UPMF, Grenoble]

Ph. D. students
Juliette Blanchet [MENRT, co-advised by F. Forbes and C. Schmid, Team Lear]
Charles Bouveyron [MENRT, co-advised by S. Girard and C. Schmid, Team Lear]
Vassil Khalidov [INRIA, co-advised by F. Forbes and S. Girard, since November 2006]
Laurent Donini [CIFRE Xerox/Inria, co-advised by S. Girard, since December 2006]
Matthieu Vignes [AC, co-advised by F. Forbes and G. Celeux, Team Select]

Post-doctoral fellows
Caroline Bernard-Michel [INRIA, December 2006-December 2007]
Chibiao Chen [INRIA, December 2005-December 2006]
Monica Benito [ERCIM, February 2006-October 2006]

Research scientists (partners)
Henri Berthelon [Faculty member, CNAM, Paris]
Gersende Fort [Research scientist, CNRS, Paris]

Administrative assistant
Claire Bonin

2. Overall Objectives

2.1. Overall Objectives

The team MISTIS aims at developing 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, ...). 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).

3. Scientific Foundations

3.1. Mixture models

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

Participants: Caroline Bernard-Michel, Juliette Blanchet, Charles Bouveyron, Florence Forbes, Gersende Fort, Stéphane Girard, Matthieu Vignes.

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 divide 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 to one of a set of K alternative categories. The distribution of an observed y_i can be written as a finite mixture of distributions,

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

These models are interesting in that they may point out an 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 parameters estimation but also values for missing data.

Mixture models correspond to independent z_i's. They are more and more used in statistical pattern recognition. They allow a formal (model-based) approach to (unsupervised) clustering.

3.2. Markov models

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

Participants: Chibiao Chen, Juliette Blanchet, Florence Forbes, Gersende Fort, Vasil Khalidov, Matthieu Vignes.

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 the 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 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.

They 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. As regards, 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

Keywords: *boundary estimation*, *dimension reduction*, *extreme value analysis*, *kernel methods*, *non parametric*.

Participants: Laurent Gardes, Stéphane Girard.

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. As an illustration, the grey-levels surface in an image cannot usually be described through a simple mathematical equation. Projection methods are then a way to decompose the unknown signal or image 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 *boundary estimation* and *image segmentation*. Such non parametric methods have become the cornerstone when dealing with functional data [47]. This is the case for instance when observations are curves. They allow to model the data without a discretization step. More generally, these techniques are of great use for dimension reduction purposes. They permit to reduce 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. This is the case in *extreme value analysis* [46], which is based on the modelling of distribution tails. It differs from traditionnal statistics which focus on the central part of distributions, *i.e.* on the most probable events. Extreme value theory shows that distributions tails can be modelled by both a functional part and a real parameter, the extreme value index. As another example, relationships exist between multiresolution analysis and parametric Markov tree models.

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 $x_1 \leq ... \leq x_n$ denote n ordered observations from a random variable X representing some quantity of interest. A p_n-quantile of X is the value q_{p_n} such that the probability that X is greater than q_{p_n} is p_n, *i.e.* $P(X > q_{p_n}) = p_n$. When $p_n < 1/n$, such a quantile is said to be extreme since it is usually greater than the maximum observation x_n (see Figure 1). To estimate such quantiles requires therefore specific methods [44], [43] to extrapolate information beyond the observed values of X. Those methods are based on Extreme value theory. This kind of issues 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. 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 [50]. These so-called Weibull-tail distributions [52], [49], [8] [21] are defined by their survival distribution function:

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

where both $\theta > 0$ and the function $\ell(x)$ are unknown. Gaussian, gamma, exponential and Weibull distributions, among others, are included in this family. The function $\ell(x)$ acts as a nuisance parameter which yields a bias in the classical extreme-value estimators developped so far.

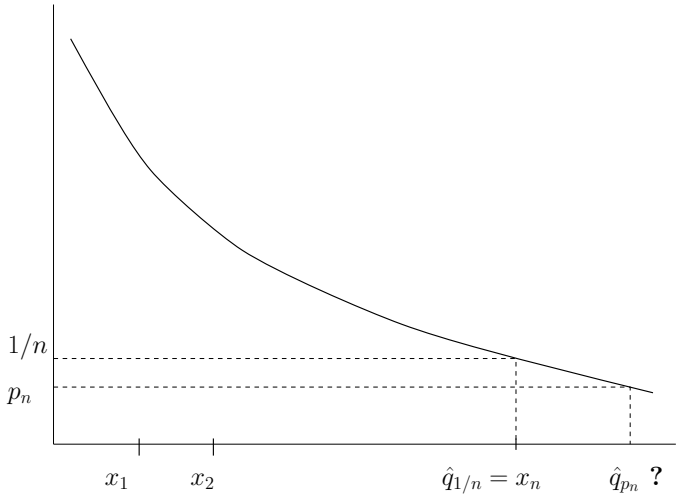


Figure 1. The curve represents the survival function $x \longrightarrow P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\hat{q}_{1/n} = x_n$. As illustrated in the figure, to estimate p_n-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.

3.3.2. Boundary estimation

Boundary estimation, or more generally, 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. In image analysis, the boundary estimation problem arises in image segmentation as well as in supervised learning.

3.3.3. Dimension reduction

Our work on high dimensional data includes non parametric aspects. They are related to Principal Component Analysis (PCA) which is traditionnaly used 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. When dealing with classification problems, our main project is then to adapt the non linear PCA method proposed in [51],[10]. This method (first introduced in Stéphane Girard's PhD thesis) relies on the approximation of datasets by manifolds, generalizing the PCA linear subspaces. This approach reveals good performances when data are images [4].

Our work also include parametric approaches in particular when considering classification and learning issues. In high dimensional spaces learning methods suffer from the curse of dimensionality: even for large datasets, large parts of the spaces are left empty. One of our approach is therefore to develop new Gaussian models of high dimensional data for parametric inference. Such models can then be used in a Mixtures or Markov framework for classification purposes.

4. Application Domains

4.1. Image Analysis

Participants: Juliette Blanchet, Charles Bouveyron, Florence Forbes, Stéphane Girard.

As regards applications, several areas of image analysis can be covered using the tools developed in the team. More specifically, we address in collaboration with Team Lear, Inria Rhone-Alpes, issues about object and class recognition and about the extraction of visual information from large image data bases.

Other applications in medical imaging are natural. We work more specifically on MRI data.

We also consider other statistical 2D fields coming from other domains such as remote sensing.

4.2. Biology and Medicine

Participants: Chibiao Chen, Florence Forbes, Stéphane Girard, Matthieu Vignes.

A second domain of applications concerns biomedical statistics and molecular biology. We consider the use of missing data models in population genetics. We also investigate statistical tools for the analysis of bacterial genomes beyond gene detection. Applications in agronomy are also considered.

4.3. Reliability

Participants: Henri Bertholon, Laurent Gardes, Stéphane Girard.

Reliability and industrial lifetime analysis are applications developed through collaborations with the EDF research department and the LCFR laboratory of CEA / Cadarache. We also consider failure detection in print infrastructure through collaborations with Xerox, Meylan.

5. Software

5.1. The HDDA and HDDC toolboxes

Participants: Charles Bouveyron, Stéphane Girard.

HDDA Toolbox. The High-Dimensional Discriminant Analysis (HDDA) toolbox contains efficient supervised classifiers for high-dimensional data. These classifiers are based on Gaussian models adapted to high-dimensional data. The HDDA toolbox is available for Matlab and will be soon included into the software MixMod. Version 1.1 of the HDDA Toolbox is now available.

HDDC Toolbox. The High-Dimensional Data Classification (HDDC) toolbox contains efficient unsupervised classifiers for high-dimensional data. These classifiers are also based on Gaussian models adapted to high-dimensional data. The HDDC toolbox is available for Matlab.

Both toolboxes are available at http://ace.acadiau.ca/math/bouveyron/softwares.html

5.2. The Extremes freeware

Participants: Laurent Gardes, Stéphane Girard.

Joint work with Jean Diebolt (CNRS), Myriam Garrido (INRA Clermont-Ferrand) and Jérôme Ecarnot.

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. 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://mistis.inrialpes.fr/software/EXTREMES/.

Recently, this software has been used to propose a new goodness-of-fit test to the distribution tail [17].

5.3. The SpaCEM^3 program

Participants: Juliette Blanchet, Florence Forbes.

The SpaCEM^3 (Spatial Clustering with EM and Markov Models) program replaces the former, still available, SEMMS (Spatial EM for Markovian Segmentation) program developed with Nathalie Peyrard from INRA Avignon.

SpaCEM^3 proposes a variety of algorithms for image segmentation, supervised and unsupervised classification of multidimensional and spatially located data. The main techniques use the EM algorithm for soft clustering and Markov Random Fields for spatial modelling. The learning and inference parts are based on recent developments based on mean field approximations. The main functionalities of the program include:

The former SEMMS functionalities, *ie.*

- Model based unsupervised image segmentation, including the following models: Hidden Markov Random Field and mixture model;
- Model selection for the Hidden Markov Random Field model;
- Simulation of commonly used Hidden Markov Random Field models (Potts models).
- Simulation of an independent Gaussian noise for the simulation of noisy images.

And additional possibilities such as,

- New Markov models including various extensions of the Potts model and triplets Markov models;
- Additional treatment of very high dimensional data using dimension reduction techniques within a classification framework;
- Models and methods allowing supervised classification with new learning and test steps.

The SEMMS package, written in C, is publicly available at: http://mistis.inrialpes.fr/software/SEMMS.html. The SpaCEM^3 written in C++ is available at http://mistis.inrialpes.fr/software/SpaCEM3.tgz.

5.4. The FASTRUCT software

Participants: Chibiao Chen, Florence Forbes.

This is joint work with Olivier Francois (TimB, TIMC).

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 [63]. 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 [15].

6. New Results

6.1. Mixture models

6.1.1. Taking into account the curse of dimensionality.

Participants: Charles Bouveyron, Stéphane Girard.

Joint work with Serge Iovleff (Université Lille 3) and Cordelia Schmid (Lear, Inria).

In the PhD work of Charles Bouveyron (co-advised by Cordelia Schmid from the INRIA team LEAR) [11], 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 new supervised classification methods called HDDA for High Dimensional Discriminant Analysis [12], [13]. 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 HDDC for High Dimensional Data Clustering [26], In collaboration with Gilles Celeux and Charles Bouveyron we are currently working on the automatic selection of the discrete parameters of the model. We also, in the context of Juliette Blanchet PhD work (also co-advised with C. Schmid), combined the method to our Markov-model based approach of learning and classification and obtained significant improvement in applications such as texture recognition [24], [25], where the observations are high-dimensional.

We are then also willing to get rid of the Gaussian assumption. To this end, non linear models and semi parametric methods are necessary.

6.1.2. Supervised and unsupervised classification of objects in images

Participants: Charles Bouveyron, Stéphane Girard.

This is joint work with Cordelia Schmid, (LEAR, INRIA Rhône-Alpes)

Supervised framework. In this framework, small scale-invariant regions are detected on a learning image set and they are then characterized by the local descriptor SIFT [61]. The object is recognized in a test image if a sufficient number of matches with the learning set is found. The recognition step is done using supervised classification methods. Frequently used methods are Linear Discriminant Analysis (LDA) and, more recently, kernel methods (SVM) [59]. In our approach, the object is represented as a set of object parts. As an example fora motorbike, we will consider three parts: wheels, seat and handlebars.

Obtained results showed that the HDDA method described in Section 6.1.1 gives better recognition results than SVM and other generative methods. In particular, the classification errors are significantly lower for HDDA compared to SVM. In addition, HDDA method is as fast as standard discriminant analysis (computation time $\simeq$ 1 sec. for 1000 descriptors) and much faster than SVM ($\simeq$ 7 sec.).

Unsupervised framework. Our approach learns automatically discriminant object parts and then identifies local descriptors belonging to the object. It first extracts a set of scale-invariant descriptors and then learns a set of discriminative object parts based on a set of positive and negative images. Learning is "weakly supervised" since objects are not segmented in the positive images. Recognition matches descriptors of a unknown image to the discriminative object parts.

Object localization is a challenging problem since it requires a very precise classification of descriptors. For this, it is necessary to identify the descriptors of an image which have a high probability to belong to the object. The adaptation of HDDA to the unsupervised framework, called HDDC, allows to compute the posterior probability for each interest point that it belongs to the object. Finally, the object can be located in a test image by considering the points with the highest probabilities. In practice, 5 or 10 percents of all detected interest points are enough to locate efficiently the object. See an illustration in Figure 2.

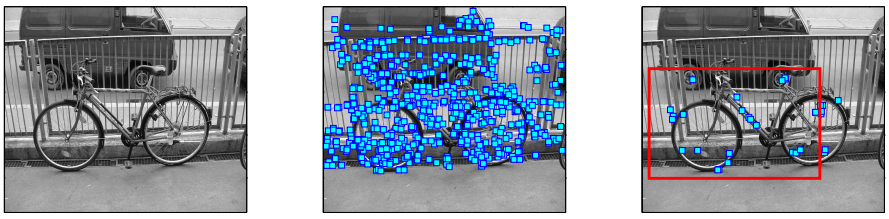


Figure 2. Object localization: from left to right, Original image, All detected points, Final localization.

We also consider the application of image classification [42]. This step decides if the object is present in the image, i.e. it classifies the image as positive (containing the object) or negative (not containing the object). We use our decision rule to assign a posterior probability to each descriptor and each cluster. We then decide based on these probabilities if a test image contains the object or not. Previous approaches [45] have used a simple empirical technique to classify a test image. We introduce a probabilistic technique which uses the posterior probabilities. We obtain for a test image I a score $S \in [0, 1]$ that I contains the object. We decide that a test image contains the object if the score S is larger than a given threshold. This probabilistic decision has the advantage of not introducing an additional parameter and of using the posterior probability to reject (assign a low weight) to dubious points [27].

6.2. Markov models

6.2.1. Supervised classification of complex structure data using mixture models and triplet Markov fields.

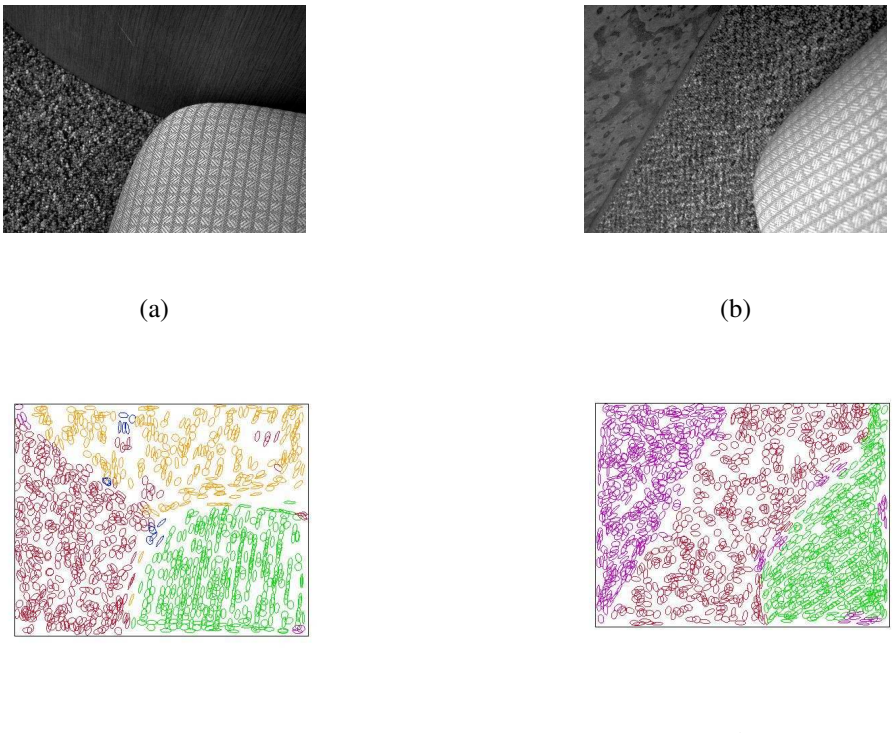
Participants: Florence Forbes, Juliette Blanchet.

In this work, we focus on three sources of complexity. We consider data exhibiting (complex) dependence structures, having to do for example with spatial or temporal association, family relationship, and so on. More specifically, we consider observations associated to sites or items at spatial locations. These locations can be irregularly spaced. This goes beyond the standard regular lattice case traditionnaly used in image analysis and requires some adaptation.

A second source of complexity is connected with the measurement process, such as having multiple measuring instruments or computations generating high dimensional data. There are not so many 1-dimensional distributions for continuous variables that generalize to multidimensional ones except when considering product of 1-dimensional independent components. The Gaussian distribution is the most commonly used but it has the specificity to be unimodal. Also, what we consider as a third source of complexity is that in real-world applications, data cannot usually be reduced to classes modeled by unimodal distributions and consequently by single gaussian distributions.

In this work, we consider supervised classification problems in which training sets are available and correspond to data for which data exemplars have been grouped into classes.

We propose a unified Markovian framework for both learning the class models and then consequently classify observed data into these classes. We show that models able to deal with the above sources of complexity can be derived based on traditional tools such as mixture models and Hidden Markov fields. For the latter, however, non trivial extensions in the spirit of [39] are required to include a learning step while preserving the Markovian modelling of the depedencies. Applications of our models include textured image segmentation. See an illustration in Figure 3.



(a) (b)

(c) (d)

Figure 3. Texture recognition using Hidden Markov Models: (a) and (b) original multi-texture images, (c) and (d) classification results (ellipses represent interest points and associated regions). The different colors correspond to different texture assignments.

6.2.2. Integrated Markov models for clustering genes

Participants: Florence Forbes, Matthieu Vignes.

Clustering of genes into groups sharing common characteristics is a useful exploratory technique for a number of subsequent computational and biological analysis. A wide range of clustering algorithms have been proposed in particular to analyze gene expression data but most of them consider genes as independent entities or include relevant information on gene interactions in a sub-optimal way.

We propose a probabilistic model that has the advantage to account for individual data (*eg.* expression) and pairwise data (*eg.* interaction information coming from biological networks) simultaneously. Our model is based on hidden Markov random field models in which parametric probability distributions account for the distribution of individual data. Data on pairs, possibly reflecting distances or similarity measures between genes, are then included through a graph where the nodes represent the genes and the edges are weighted according to the available interaction information. As a probabilistic model, this model has many interesting theoretical features. Also, preliminary experiments on simulated and real data show promising results and points out the gain in using such an approach [34], [35], [37], [36].

6.2.3. Distributed and Cooperative Markovian segmentation of both tissues and structures in brain MRI.

Participant: Florence Forbes.

This is joint work with Benoit Scherrer, Michel Dojat and Christine Garbay from INSERM and LIG.

Accurate tissue and structure segmentation of MRI brain scan is critical for several applications. Markov random fields are commonly used for tissue segmentation to take into account spatial dependencies between voxels, hence acting as a labelling regularization. However, such a task requires the estimation of the model parameters (eg. Potts model) which is not tractable without approximations. The algorithms in [3] based on EM and variational approximations are considered. They show interesting results for tissue segmentation but are not sufficient for structure segmentation without introducing a priori anatomical knowledge. In most approaches, structure segmentation is performed after tissue segmentation. We suggest considering them as combined processes that cooperate. Brain anatomy is described by fuzzy spatial relations between structures that express general relative distances, orientations or symmetries. This knowledge is incorporated into a 2-class Markov model via an external field. This model is used for structure segmentation. The resulting structure information is then incorporated in turn into a 3 to 5-class Markov model for tissue segmentation via another specific external field. Tissue and structure segmentations thus appear as dynamical and cooperative MRF procedures whose performance increases gradually. This approach is implemented into a multi-agent framework, where autonomous entities, distributed into the image, estimate local Markov fields and cooperate to ensure consistency [32], [33]. We show, using phantoms and real images (acquired on a 3T scanner), that a distributed and cooperative Markov modelling using anatomical knowledge is a powerful approach for MRI brain scan segmentation (See Figure 4).

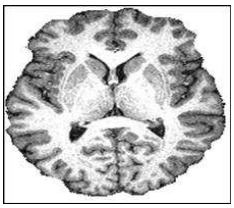
The current investigation concerns only one type (T1) of MR images with no temporal information. We are planning to extend our tools to include multidimensional MR sequences corresponding to other types of MR modalities and longitudinal data.

6.2.4. Modelling and inference of population structure from genetic and spatial data

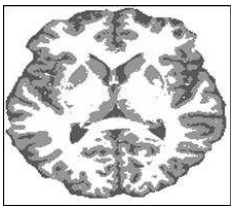
Participants: Chibiao Chen, Florence Forbes.

This is joint work with Olivier François from team TimB in TIMC laboratory.

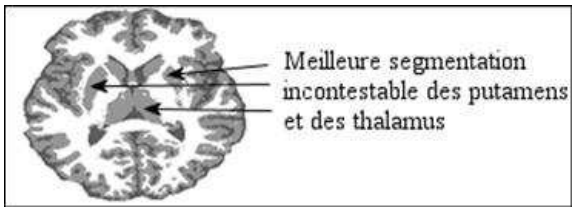
In applications of population genetics, it is often useful to classify individuals in a sample into populations which become then the units of interest. However, the definition of populations is typically subjective, based, for example, on linguistic, cultural, or physical characters as well as the geographic location of sampled individuals. Recently, Pritchard et al [63], proposed a Bayesian approach to classify individuals into groups using genotype data. Such data, also called multilocus genotype data, consists of several genetic markers



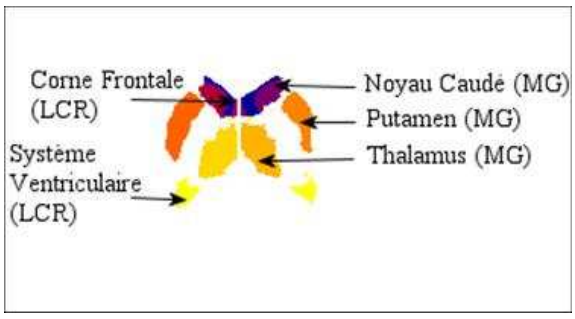
(a)



(b)



(c)



(d)

Figure 4. Distributed and cooperative Markovian segmentation: (a) real 3T scan, (b) tissue segmentation without anatomical knowledge, (c) and (d) tissue and structure segmentations using the distributed and cooperative approach.

whose variations are measured at a series of loci for each sampled individual. Their method is based on a parametric model (model-based clustering) in which there are K groups (where K may be unknown), each of which is characterized by a set of allele frequencies at each locus. Group allele frequencies are unknown and modeled by a Dirichlet distribution at each locus within each group. A MCMC algorithm is then used to estimate simultaneously assignment probabilities and allele frequencies for all groups. In such a model, individuals are assumed to be independent, which does not take into account their possible spatial proximity.

The main goal of this work is to introduce spatial prior models and to assess their role in accounting for the relationships between individuals. In this perspective, we propose to investigate particular Markov models on graphs and to evaluate the quality of mean field approximations for the estimation of their parameters.

Maximum likelihood estimation of such models in a spatial context is typically intractable but mean field like approximations within an EM algorithm framework, in the spirit of [3] will be considered to deal with this problem. This should result in a procedure alternative to MCMC approaches. With this in mind, we first considered the EM approach in a non spatial case, as an alternative to the traditional Bayesian approaches. The corresponding new computer program (see Section 5.4) and promising results are reported in [15].

6.2.5. Statistical methods for the visualization and analysis of complex remote sensing data

Participants: Caroline Bernard-Michel, Juliette Blanchet, Florence Forbes, Laurent Gardes, Stéphane Girard.

This is joint work with Sylvain Douté and Etienne Deforas from Laboratoire de Planétologie de Grenoble, France.

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 the 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.

A new generation of imaging spectrometers is emerging with an additional angular dimension. The surface of the planets will now be observed from different view points on the satellite trajectory, corresponding to about ten different angles, instead of only one corresponding usually to the vertical (0 degree angle) view point. Multi-angle imaging spectrometers present several advantages: the influence of the atmosphere on the signal can be better identified and separated from the surface signal on focus, the shape and size of the surface components and the surfaces granularity can be better characterized.

However, this new generation of spectrometers also results in a significant increase in the size (several terabits expected) and complexity of the generated data. Consequently, HMA (Hyperspectral Multi Angular) data induce data manipulation and visualization problems due to its size and its 4 dimensionality.

We propose to investigate the use of statistical techniques to deal with these generic sources of complexity in data beyond the traditional tools in mainstream statistical packages. Our goal is twofold:

- we first focus on developing or adapting dimension reduction methods, classification and segmentation methods for informative, useful visualization and representation of the data previous to its subsequent analysis.
- We also address the problem of physical model inversion which is important to understand the complex underlying physics of the HMA signal formation. The models taking into account the angular dimension result in more complex treatments. We investigate the use of semiparametric dimension reduction methods such as SIR (Sliced Inverse Regression, [60]) to perform model inversion, in a reasonable computing time, when the number of input observations increases considerably.

The first data set under consideration (hyperspectral images with vertical pointing) comes from the Mars-Express Mission operated by the European Space Agency. The second data set (multi-angular hyperspectral images) will be generated by the CRISM instrument of the Mars Reconnaissance Orbiter (NASA) that has

started its scientific activities in June 2006 after orbit insertion. LPG is a co-investigator of the CRISM instrument.

6.3. Semi and non parametric methods

6.3.1. Modelling extremal events

Participants: Stéphane Girard, Laurent Gardes.

This is joint work with Cécile Amblard (TimB in TIMC laboratory, Univ. Grenoble 1), Myriam Garrido (INRA Clermont-Ferrand), Armelle Guillou (Univ. Strasbourg), and Jean Diebolt (CNRS, Univ. Marne-la-vallée).

Our first achievement is the development of new estimators: kernel estimators and bias correction through exponential regression [16]. Our second achievement is the construction of a goodness-of-fit test for the distribution tail. Usual tests are not adapted to this problem since they essentially check the adequation to the central part of the distribution. Next, we aim at adapting extreme-value estimators to take into account covariate information. Such estimators would include extreme conditional quantiles estimators, which are closely linked to the frontier estimators presented in Section 6.3.2. Finally, more future work will include the study of multivariate extreme values. To this aim, a research on some particular copulas [1], [38] has been initiated with Cécile Amblard, since they are the key tool for building multivariate distributions [62].

6.3.2. Boundary estimation

Participants: Stéphane Girard, Laurent Gardes.

This is joint work with Anatoli Iouditski (Univ. Joseph Fourier, Grenoble), Guillaume Bouchard (Xerox, Meylan), Pierre Jacob and Ludovic Menneteau (Univ. Montpellier 2) and Alexandre Nazin (IPU, Moscow, Russia).

Two different and complementary approaches are developped.

6.3.2.1. Extreme quantiles approach.

Here, 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 propose estimators based on projection as well as on kernel regression methods applied on the extreme values set [57], [56], [55], [58], for particular set of points. In this framework, we can obtain the asymptotic distribution of the error between estimators and the true frontier [53],[22]. Our future work will be to define similar methods based on wavelets expansions in order to estimate non-smooth boundaries, and on local polynomials estimators to get rid of boundary effects. Besides, we are also working on the extension of our results to more general sets of points. This work has been initiated in the PhD work of Laurent Gardes [48], co-directed by Pierre Jacob and Stéphane Girard and in [31] with the consideration of star-shaped supports.

6.3.2.2. Linear programming approach.

Here, the boundary of a set of points is defined has a closed curve bounding all the points and with smallest associate surface. It is thus natural to reformulate the boundary estimation method as a linear programming problem [41], [40], [54]. The resulting estimate is parsimonious, it only relies on a small number of points. This method belongs to the Support Vector Machines (SVM) techniques. Their finite sample performances are very impressive but their asymptotic properties are not very well known, the difficulty being that there is no explicit formula of the estimator. However, such properties are of great interest, in particular to reduce the estimator bias. Two directions of research will be investigated. The first one consists in modifying the optimization problem itself. The second one is to use *Jacknife* like methods, combining two biased estimators so that the two bias cancel out. One of the goals of our work is also to establish the speed of convergence of such methods in order to try to improve them.

6.3.3. Modelling nuclear plants

Participants: Laurent Gardes, Stéphane Girard.

This is joint work with Nadia Perot, Nicolas Devictor and Michel Marquès (CEA).

One of the main activities of the Laboratoire de Conduite et Fiabilité des Réacteurs (CEA Cadarache) concerns the probabilistic analysis of some processes using reliability and statistical methods. In this context, probabilistic modelling of steels tenacity in nuclear plants tanks has been developed. The databases under consideration include hundreds of data indexed by temperature, so that, reliable probabilistic models have been obtained for the central part of the distribution.

However, in this reliability problem, the key point is to investigate the behaviour of the model in the distribution tail. In particular, we are mainly interested in studying the lowest tenacities when the temperature varies. We are currently investigating the opportunity to propose a postdoctoral position on this problem, supported by the CEA.

7. Contracts and Grants with Industry

7.1. Contracts

We signed in december 2006 a CIFRE contract with Xerox, Meylan, regarding the PhD work of Laurent Donini about statistical techniques for mining logs and usage data in a print infrastructure. The thesis will be co-advised by Stéphane Girard and Jean-Michel Renders (Xerox).

8. Other Grants and Activities

8.1. 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.

8.2. National initiatives

MISTIS got a Ministry grant (Action Concertée Incitative Masses de données) for a three-year project involving other partners (Team Lear from INRIA, SMS from University Joseph Fourier and Heudiasyc from UTC, Compiègne). The project called Movistar aims at investigating visual and statistical models for image recognition and description and learning techniques for the management of large image databases.

Since July 2005, MISTIS is also involved in the IBN (Integrated Biological Networks) project coordinated by Marie-France Sagot from INRIA team HELIX. This project is part of the Cooperative Research Initiative (ARC) supported by INRIA. The other partners include two other INRIA teams (HELIX and SYMBIOSE, Pasteur Institute and INRA, Jouy-en-Josas.

8.3. International initiatives

8.3.1. Europe

J. Blanchet, C. Bouveyron, F. Forbes and S. Girard are members of the Pascal Network of Excellence.

S. Girard is a member of the European project (Interuniversity Attraction Pole network) “Statistical techniques and modelling for complex substantive questions with complex data”,
Web site : http://www.stat.ucl.ac.be/IAP/frameiap.html.

S. Girard has also joint work with Prof. A. Nazin (Institute of Control Science, Moscow, Russia).

MISTIS is then involved in a European STREP proposal, named POP (Perception On Purpose) coordinated by Radu Horaud from INRIA team MOVI. The three-year project starts in January 2006. Its objective is to put forward the modelling of perception (visual and auditory) as a complex attentional mechanism that embodies a decision taking process. The task of the latter is to find a trade-off between the reliability of the sensorial stimuli (bottom-up attention) and the plausibility of prior knowledge (top-down attention). The MISTIS part and in particular the PhD work of Vasil Kalidhov is to contribute to the development of theoretical and algorithmic models based on probabilistic and statistical modelling of both the input and the processed data. Bayesian theory and hidden Markov models in particular will be combined with efficient optimization techniques in order to confront physical inputs and prior knowledge.

8.3.2. North Africa

S. Girard has joint work with M. El Aroui (ISG Tunis).

8.3.3. North America

F. Forbes has joint work with:

- C. Fraley (Univ. of Washington, USA)

- A. Raftery (Univ. of Washington, USA)

9. Dissemination

9.1. Leadership within scientific community

F. Forbes is member of the group in charge of incentive initiatives (GTAI) in the Scientific and Technological Orientation Council (COST) of INRIA.

S. Girard was involved in the PhD commitee of Charles Bouveyron from university Joseph Fourier. Title of the thesis in French: Modelisation et classification de données de grandes dimensions, application à l'analyse d'images.

He was also involved in the PhD commitee of Aurélie Muller from university Montpellier 2. Title of the thesis: Comportement asymptotique de la distribution des pluies extremes en France.

9.2. University Teaching

F. Forbes lectured a graduate course on the EM algorithm at Univ. J. Fourier, Grenoble.

L. Gardes is faculty member at Univ. P. Mendes France and Stéphane Girard was faculty member at Univ. J. Fourier in Grenoble until june 2006.

H. Berthelon is faculty member at CNAM, Paris.

9.3. Conference and workshop committees, invited conferences

Florence Forbes and Matthieu Vignes were invited to the 31st conference on Stochastic Processes and their Applications in Paris, France.

Stéphane Girard was invited speaker at the workshop on Principal manifolds for data cartography and dimension reduction in Leceister, UK, august 2006 and at the IMS annual meeting and X Brazilian School of Probability in Rio de Janeiro, Brazil, July 2006.

10. Bibliography

Major publications by the team in recent years

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

[2] G. CELEUX, S. CHRÉTIEN, F. FORBES, A. MKHADRI. *A Component-wise EM Algorithm for Mixtures*, in "Journal of Computational and Graphical Statistics", vol. 10, 2001, p. 699–712.

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

[4] B. CHALMOND, S. GIRARD. *Nonlinear modeling of scattered multivariate data and its application to shape change*, in "IEEE Trans. PAMI", vol. 21(5), 1999, p. 422–432.

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

[6] F. FORBES, A. E. RAFTERY. *Bayesian Morphology: Fast Unsupervised Bayesian Image analysis*, in "Journal of the American Statistical Association", vol. 94, n^{o} 446, 1999, p. 555-568.

[7] G. FORT, E. MOULINES. *Convergence of the Monte-Carlo EM for curved exponential families*, in "Annals of Statistics", vol. 31, n^{o} 4, 2003, p. 1220-1259.

[8] L. GARDES, S. GIRARD. *Estimating extreme quantiles of Weibull tail-distributions*, in "Communication in Statistics - Theory and Methods", vol. 34, 2005, p. 1065–1080.

[9] S. GIRARD. *A nonlinear PCA based on manifold approximation*, in "Computational Statistics", vol. 15(2), 2000, p. 145-167.

[10] S. GIRARD, S. IOVLEFF. *Auto-Associative Models and Generalized Principal Component Analysis*, in "Journal of Multivariate Analysis", vol. 93, n^{o} 1, 2005, p. 21–39.

Year Publications

Doctoral dissertations and Habilitation theses

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

Articles in refereed journals and book chapters

[12] C. BOUVEYRON, S. GIRARD, C. SCHMID. *Class-specific subspace discriminant analysis for high-dimensional data*, in "Lecture Notes in Computer Science, Berlin Heidelberg", C. SAUNDER (editor). , vol. 3940, Springer-Verlag, 2006, p. 139–150.

[13] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High Dimensional Discriminant analysis*, in "Communications in Statistics", to appear, 2006.

[14] G. CELEUX, F. FORBES, C. ROBERT, M. TITTERINGTON. *Deviance Information Criteria for missing data models. With discussion*, in "Bayesian Analysis", to appear, 2006.

[15] C. CHEN, F. FORBES, O. FRANCOIS. *FASTRUCT: Model-based clustering made faster*, in "Molecular Ecology Notes", to appear, 2006.

[16] J. DIEBOLT, L. GARDES, S. GIRARD, A. GUILLOU. *Bias-reduced estimators of the Weibull tail-coefficient*, in "Test", to appear, 2006.

[17] J. DIEBOLT, M. GARRIDO, S. GIRARD. *A Goodness-of-fit Test for the Distribution Tail*, in "Topics in extreme values, New-York", M. AHSANULLAH, S. KIRMANI (editors). , to appear, Nova Science, 2006.

[18] F. FORBES, G. FORT. *Combining Monte Carlo and Mean field like methods for inference in hidden Markov Random Fields*, in "IEEE trans. on Image Processing", to appear, 2006.

[19] F. FORBES, N. PEYRARD, C. FRALEY, D. GEORGIAN-SMITH, D. GOLDHABER, A. RAFTERY. *Model-Based Region-of-Interest Selection in Dynamic Breast MRI*, in "Journal of Computer Assisted Tomography", vol. 30, n^{o} 4, July/August 2006, p. 675-687.

[20] L. GARDES, S. GIRARD. *Asymptotic properties of a Pickands type estimator of the extreme value index*, in "Focus on probability theory, New-York", L. R. VELLE (editor). , Nova Science, 2006, p. 133–149.

[21] L. GARDES, S. GIRARD. *Comparison of Weibull tail-coefficient estimators*, in "REVSTAT - Statistical Journal", vol. 4, n^{o} 2, 2006, p. 373–188.

[22] J. GEFFROY, S. GIRARD, P. JACOB. *Asymptotic normality of the L_1-error of a boundary estimator*, in "Nonparametric Statistics", vol. 18, n^{o} 1, 2006, p. 21–31.

Publications in Conferences and Workshops

[23] C. AMBLARD, S. GIRARD. *A semiparametric family of bivariate copulas: dependence properties and estimation procedures*, in "IMS Annual Meeting and X Brazilian School of Probability, Rio de Janeiro, Brésil", juillet 2006.

[24] J. BLANCHET, C. BOUVEYRON. *Modèle markovien caché pour la classification supervisée de données de grande dimension spatialement corrélées*, in "38èmes Journées de Statistique de la Société Française de Statistique, Clamart, France", Mai 2006.

[25] J. BLANCHET, F. FORBES. *Triplet Markov fields designed for supervised classification of textured images*, in "COMPSTAT, 17th symposium of the IASC, Roma, Italy", 2006.

[26] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional data clustering*, in "COMPSTAT, 17th symposium of the IASC, Roma, Italy", August 2006.

[27] C. BOUVEYRON, J. KANNALA, C. SCHMID, S. GIRARD. *Object localization by subspace clustering of local descriptors*, in "5th Indian Conference on Computer Vision, Graphics and Image Processing, Madurai, Inde", décembre 2006.

[28] G. DEWAELE, F. DEVERNAY, R. HORAUD, F. FORBES. *The alignment between 3D-data and articulated shapes with bending surfaces*, European Conf. Computer Vision, 2006.

[29] S. GIRARD, A. IOUDITSKI, A. NAZIN. *On optimal and adaptive non-parametric estimation for periodic frontier via linear programming*, in "Third International Control Conference, Moscou, Russia", juin 2006.

[30] S. GIRARD, S. IOVLEFF. *Auto-Associative models and generalized Principal Component Analysis*, in "Workshop principal manifolds for data cartography and dimension reduction, Leicester, UK", aout 2006.

[31] S. GIRARD, L. MENNETEAU. *Estimation of star-shaped supports via smoothed extreme value estimators of non-uniform point processes boundaries*, in "IMS Annual Meeting and X Brazilian School of Probability, Rio de Janeiro, Brésil", juillet 2006.

[32] B. SCHRERRER, M. DOJAT, F. FORBES, C. GARBAY. *Distributed and Cooperative Markovian Segmentation of tissues and structures in MRI brain scans*, in "HBM meeting, Florence Italy", 2006.

[33] B. SCHRERRER, M. DOJAT, F. FORBES, C. GARBAY. *Segmentation Markovienne distribuée et coopérative des tissus et des structures présents dans des IRM cérébrales*, in "RFIA, Tours, France", 2006.

[34] M. VIGNES, F. FORBES. *A statistical glance at clustering models to fit biological network and expression data*, in "31st conference on Stochastic Processes and their Applications, Paris, France", 2006.

[35] M. VIGNES, F. FORBES. *Integrated Markov models for clustering gene expression data*, in "Journées MAS de la SMAI, Lille, France", 2006.

[36] M. VIGNES, F. FORBES. *Integrated Markov models for clustering genes combining individual features and pairwise relationships*, in "4th workshop on Statistical methods for post-genomic data, Toulouse, France", 2006.

[37] M. VIGNES, F. FORBES. *Markov Random Fields for clustering genes*, in "2eme Recontres Inter-Associations: la classification et ses applications, Lyon, France", 2006.

References in notes

[38] C. AMBLARD, S. GIRARD. *Symmetry and dependence properties within a semiparametric family of bivariate copulas*, in "Nonparametric Statistics", vol. 14, n^{o} 6, 2002, p. 715–727.

[39] D. BENBOUDJEMA, W. PIECZYNSKI. *Unsupervised image segmentation using triplet Markov fields*, in "Computer Vision and Image Understanding", vol. 99, n^{o} 3, 2005, p. 476–498.

[40] G. BOUCHARD, S. GIRARD, A. IOUDITSKI, A. NAZIN. *Nonparametric Frontier estimation by linear programming*, in "Automation and Remote Control", vol. 65, n^{o} 1, 2004, p. 58–64.

[41] G. BOUCHARD, S. GIRARD, A. IOUDITSKI, A. NAZIN. *Some Linear programming methods for frontier estimation*, in "Applied Stochastic Models in Business and Industry", vol. 21, n^{o} 2, 2005, p. 175–185.

[42] C. BOUVEYRON, S. GIRARD, C. SCHMID. *Une nouvelle méthode de classification pour la reconnaissance de formes*, in "20e colloque GRETSI sur le traitement du signal et des images, Louvain-la-Neuve, Belgium", September 2005.

[43] J. DIEBOLT, M. EL-AROUI, M. GARRIDO, S. GIRARD. *Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling*, in "Extremes", vol. 8, 2005, p. 57–78.

[44] J. DIEBOLT, S. GIRARD. *A Note on the asymptotic normality of the ET method for extreme quantile estimation*, in "Statistics and Probability Letters", vol. 62, n^{o} 4, 2003, p. 397–406.

[45] G. DORKÓ, C. SCHMID. *Object Class Recognition Using Discriminative Local Features*, Submitted to IEEE Trans. on Pattern Analysis and Machine Intelligence, updated 13 September, 2005.

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

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

[48] L. GARDES. *Estimation d'une fonction quantile extrême*, Ph. D. Thesis, Université Montpellier 2, october 2003.

[49] L. GARDES, S. GIRARD. *Estimating extreme quantiles of Weibull-tail distributions*, in "STATDEP, Statistics for dependent data, Paris-Malakoff", janvier 2005.

[50] L. GARDES, S. GIRARD. *Statistical Inference for Weibull-tail distributions*, in "Workshop on risk analysis and extreme values, Paris", juin 2005.

[51] S. GIRARD. *A nonlinear PCA based on manifold approximation*, in "Computational Statistics", vol. 15(2), 2000, p. 145–167.

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

[53] S. GIRARD. *On the asymptotic normality of the L_1- error for Haar series estimates of Poisson point processes boundaries*, in "Statistics and Probability Letters", vol. 66, 2004, p. 81–90.

[54] S. GIRARD, A. IOUDITSKI, A. NAZIN. *L_1-Optimal Nonparametric Frontier Estimation via Linear Programming*, in "Automation and Remote Control", vol. 66, n^{o} 12, 2005, p. 2000–2018.

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

[56] S. GIRARD, P. JACOB. *Projection estimates of point processes boundaries*, in "Journal of Statistical Planning and Inference", vol. 116, n^{o} 1, 2003, p. 1–15.

[57] S. GIRARD, P. JACOB. *Extreme values and kernel estimates of point processes boundaries*, in "ESAIM: Probability and Statistics", vol. 8, 2004, p. 150–168.

[58] S. GIRARD, L. MENNETEAU. *Central limit theorems for smoothed extreme value estimates of point processes boundaries*, in "Journal of Statistical Planning and Inference", vol. 135, n^{o} 2, 2005, p. 433-460.

[59] T. HASTIE, R. TIBSHIRANI, J. FRIEDMAN. *The Elements of Statistical Learning*, Springer, NewYork, 2001.

[60] K. LI. *Sliced inverse regression for dimension reduction*, in "Journal of the American Statistical Association", vol. 86, 1991, p. 316–342.

[61] D. LOWE. *Distinctive image features from scale-invariant keypoints*, in "International Journal of Computer Vision", vol. 60, n^{o} 2, 2004, p. 91-110.

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

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