INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team MISTIS

Modelling and Inference of Complex and Structured Stochastic Systems

Grenoble - Rhône-Alpes

THEME COG

Table of contents

1. Team

Team leader
Florence Forbes [Research scientist, INRIA]

Administrative assistant
Barta Angles

Research scientists
Laurent Gardes [Faculty member, UPMF, Grenoble]
Stéphane Girard [Research scientist, INRIA, HdR]
Jean-Baptiste Durand [60%, Faculty member, INPG, Grenoble]

External collaborators
Gersende Fort [20%, Research scientist, CNRS, Paris]

Ph. D. students
Juliette Blanchet [MENRT, co-advised by F. Forbes and C. Schmid, Team Lear, until October 2007]
Laurent Donini [CIFRE Xerox/INRIA, co-advised by J.B. Durand and S. Girard]
Vasil Khalidov [INRIA, co-advised by F. Forbes and S. Girard]
Alexandre Lekina [INRIA, co-advised by L. Gardes and S. Girard, since september 2007]
Matthieu Vignes [AC, co-advised by F. Forbes and G. Celeux, Team Select, until October 2007]

Post-doctoral fellows
Caroline Bernard-Michel [INRIA, December 2006-December 2007]
Senan Doyle [INRIA, since December 2007]

Student interns
Rajendran Narayanan [INRIA, June-July 2007]

Technical staff
Sophie Chopart [Research Engineer, since September 2007]

2. Overall Objectives

2.1. Introduction

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).

2.2. Highlights of the year

MISTIS got Ministry grants for two interdisciplinary ANR projects. The first one is called "Visualisation et analyse d'images hyperspectrales multidimensionnelles en Astrophysique" (VAHINEES). and aims at developing physical as well as mathematical models, algorithms, and software able to deal efficiently with hyperspectral multi-angle data but also with any other kind of large hyperspectral dataset (astronomical or experimental). The second one is called "Forecast and projection in climate scenario of Mediterranean intense events: Uncertainties and Propagation on environment" (MEDUP) and deals with the quantification and identification of sources of uncertainties associated with the forecast and climate projection for Mediterranean high-impact weather events.

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: Juliette Blanchet, Jean-Baptiste Durand, 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,

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

These models are interesting in that they may point out 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: Juliette Blanchet, Jean-Baptiste Durand, 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: *dimension reduction*, *extreme value analysis*, *kernel method*, *level sets estimation*, *non-parametric*, *projection methods*.

Participants: Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard, Alexandre Lekina.

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 paragraph 3.3.2. Such non-parametric methods have become the cornerstone when dealing with functional data [39]. 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 (paragraph 3.3.3). 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. Examples include the Sliced Inverse Regression (SIR) method [44] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [38], which is based on the modelling of distribution tails, see paragraph 3.3.1. 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.

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).

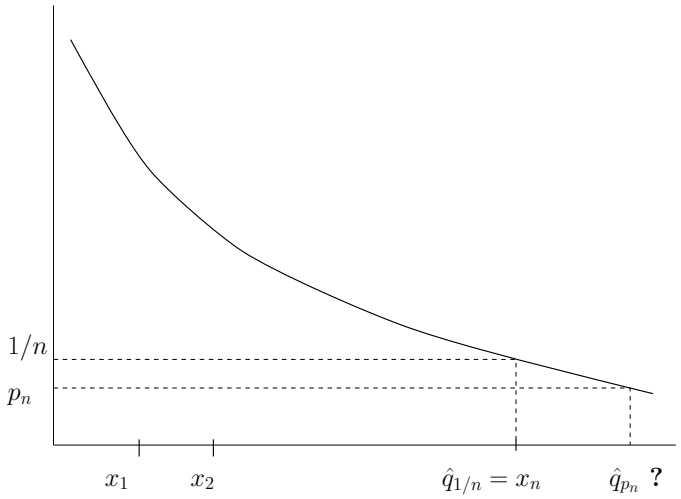


Figure 1. The curve represents the survival function $x \to P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\widehat{q}_{1/n} = x_n$. As illustrated in the figure, to estimate p_n-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.

To estimate such quantiles requires therefore dedicated methods 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. To this end, semi-parametric models of the tail are considered:

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

where both the extreme-value index $\theta > 0$ and the function $\ell(x)$ are unknown. The function $\ell(x)$ acts as a nuisance parameter which yields a bias in the classical extreme-value estimators developped so far. Such models are often refered to as heavy-tail models since the probability of extreme events decreases at a polynomial rate to zero. 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 [8] are defined by their survival distribution function:

$$P(X > x) = \exp\{-x^{\theta}\ell(x)\},\ x > x_0 > 0. \tag{3}$$

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 (3) in order to propose new estimation methods.

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 permits to benefit 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.

3.3.3. Dimension reduction

Our work on high dimensional data imposes to face the curse of dimensionality phenomenon. Indeed, the modelling of high dimensional data requires complex models and thus the estimation of high number of parameters compared to the sample size. In this framework, dimension reduction methods aim at replacing the original variables by a small number of linear combinations with as small as possible loss of information. Principal Component Analysis (PCA) is the most widely used method to reduce dimension in data. However, standard linear PCA can be quite inefficient on image data where even simple image 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 [43]. 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 [36]. Such models can then be used in a Mixtures or Markov framework for classification purposes. Another approaches consists in combining dimension reduction, regularization techniques and regression techniques to improve the Sliced Inverse Regression method [44].

4. Application Domains

4.1. Image Analysis

Participants: Caroline Bernard-Michel, Juliette Blanchet, Florence Forbes, Laurent Gardes, 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 Rhône-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. Finally, in the context of the ANR MDCO project, see paragraph 8.2, we work on hyperspectral multi-angle images.

4.2. Biology, Environment and Medicine

Participants: Florence Forbes, Laurent Gardes, Stéphane Girard, Vasil Khalidov, Alexandre Lekina, 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. Finally, in the context of the ANR VMC project, see paragraph 8.2, we plan to study the uncertainties on the forecasting and climate projection for Mediterranean high-impact weather events.

4.3. Reliability

Participants: Laurent Donini, Jean-Baptiste Durand, Laurent Gardes, 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 through collaborations with Xerox, Meylan and the CIFRE PhD thesis of Laurent Donini, co-advised by Jean-Baptiste Durand and Stéphane Girard.

5. Software

5.1. The HDDA and HDDC toolboxes

Participant: Stéphane Girard.

Joint work with: Charles Bouveyron (Université Paris 1) and Gilles Celeux (Select, INRIA). 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 [36]. The HDDA and HDDC toolboxes are available for Matlab and will be soon included into the software MixMod.

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

5.2. The Extremes freeware

Participants: Sophie Chopart, Laurent Gardes, Stéphane Girard.

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

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 [42]. 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 [16]. Besides, a new interface is going to be developped by Sophie Chopart in C++ in order to obtain a software independent of Matlab.

5.3. The SpaCEM^3 program

Participants: Juliette Blanchet, Sophie Chopart, 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. Sophie Chopart started working on an improved version including a user interface that should be available in 2008.

5.4. The FASTRUCT software

Participant: Florence Forbes.

Joint work with: Olivier Francois (TimB, TIMC) and Chibiao Chen (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 [46]. 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 [37].

5.5. The TESS software

Participant: Florence Forbes.

Joint work with: Olivier Francois (TimB, TIMC) and Chibiao Chen (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 [40]. 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 from 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 [13].

6. New Results

6.1. Mixture models

6.1.1. Taking into account the curse of dimensionality.

Participant: Stéphane Girard.

Joint work with: Charles Bouveyron (Université Paris 1), Gilles Celeux (Select, INRIA) and Cordelia Schmid (Lear, INRIA).

In the PhD work of Charles Bouveyron (co-advised by Cordelia Schmid from the INRIA team LEAR) [36], 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 HDDA for High Dimensional Discriminant Analysis [11]. 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 [10]. 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 by 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 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. Multi-speaker Localization with Binaural Audition and Stereo Vision using the EM Algorithm

Participants: Florence Forbes, Vasil Khalidov.

Joint work with: Elise Arnaud, Miles Hansard, Radu Horaud and Ramya Narasimha from the INRIA team Perception.

This work takes place in the context of the POP European project (see Section 8.3.1) and includes further collaborations with researchers from University of Sheffield, UK. The context is that of multi-modal sensory signal integration. We focus on audio-visual integration. Fusing information from audio and video sources has resulted in improved performance in applications such as tracking. However, crossmodal integration is not trivial and requires some cognitive modelling because at a lower level, there is no obvious way to associate depth and sound sources. Combining expertise from team Perception and University of Sheffield, we address the difficult problems of integrating spatial and temporal audio-visual stimuli using a geometrical and probabilistic framework and attack the problem of associating sensorial descriptions with representation of prior knowledge.

First, we address the problem of speaker localization within an unsupervised model-based clustering framework. Both auditory and visual observations are available. We gather observations over a time interval $[t_1, t_2]$. We assume that within this time interval the speakers are static so that each speaker can be described by its 3-D location in space. A cluster is associated with each speaker. In practice we consider $N + 1$ possible clusters corresponding to the addition of an extra outlier category to the N speakers.

We then consider then a set of M visual observations. Each such observation corresponds to a binocular disparity, namely a 3-D vector $\mathbf{Y}_m = (u_m, v_m, d_m)^t$ where u_m and v_m correspond to the 2-D location in the Cyclopean image[1], and d_m denotes the measured disparity at this image location. Note that such a binocular disparity corresponds to the location of a physical object that is visible in both the left and right images of the stereo pair. We define a function $v : \mathcal{R}^3 \rightarrow \mathcal{R}^3$ such that $v(\mathbf{s}_n)$ represents the binocular disparity of speaker n when his location is given by $\mathbf{s}_n$.

[1] The Cyclopean image is a geometric construction developped by M. Hansard and R. Horaud

Similarly, let us consider a set of K auditory observations. Each such observation corresponds to an auditory disparity, namely the *interaural time difference*, or ITD. We define a function $u : \mathcal{R}^3 \rightarrow \mathcal{R}$ such that $u(\mathbf{s}_n)$ evaluates the ITD of speaker n given his coordinates in the 3-D space.

We then show that recovering speakers localizations can be seen as a parameter estimation issue in a missing data framework. The parameters to be estimated are the speaker locations, and the missing variables are the assignement variables associating each individual observations to one of the N speakers ot to the outlier class. We are currently investigating the use of the EM algorithm to provide these parameters estimates.

6.2. Markov models

6.2.1. Triplet Markov fields for the classification of complex structure data

Participants: Florence Forbes, Juliette Blanchet.

We address the issue of classifying complex data. We focus on three main sources of complexity, namely the high dimensionality of the observed data, the dependencies between these observations and the general nature of the noise model underlying their distribution. We investigate the recent *Triplet Markov Fields* and propose [9] new models in this class designed for such data and in particular allowing very general noise models. In addition, our models can handle the inclusion of a learning step in a consistent way so that they can be used in a supervised framework. One other advantage of our models is that whatever the initial complexity of the noise model, parameter estimation can be carried out using state-of-the-art Bayesian clustering techniques under the usual simplifying assumptions (typically, non correlated noise condition). As generative models, they can be seen as an alternative, in the supervised case, to discriminative Conditional Random Fields. In the non supervised case, identifiability issues underlying the models can occur. We also consider the issue of selecting the best model with regards to the observed data using a criterion (referred to as BIC^{MF}) based on the Bayesian Information Criterion (BIC).

In [9], the models performance is illustrated on simulated and real data exhibiting the mentioned various sources of complexity. See also Figure 2 for an illustration on synthetic data.

6.2.2. Integrated Markov models for clustering genes: combining expression data with missing values and gene interaction network analysis

Participants: Juliette Blanchet, Florence Forbes, Matthieu Vignes.

DNA microarray technologies provide means for monitoring in the order of tens of thousands of gene expression levels quantitatively and simultaneously. However data generated in these experiments can be noisy and have missing values. When it is not ignored, the last issue has been solved by imputing the expression matrix in order to keep going with traditional analysis method. Although it was a first useful step, it is not recommended to use value imputation to deal with missing data. Moreover, appropriate tools are needed to cope with noisy background in expression levels and to take into account a dependency structure among genes under study. Various approaches have been proposed but to our knowledge none of them has the ability to fulfil all these features. We therefore propose [26] a clustering algorithm that explicitly accounts for dependencies within a biological network and for missing value mechanism to analyze microarray data. We propose to tackle these issues in a unique statistical framework. We take advantage of many features of the probabilistic aspect of the model. In a previous work [22], we mentioned the ability of a straightforward extension of the model therein to deal with missing values. It is now implemented and we prove it to be successful at dealing with different absence patterns either on simulated or real biological data sets. We emphasize that our model can be useful in a great range of applications for clustering entities of interest (such as genes, proteins, metabolites in post-genomics studies). It requires individual possibly incomplete measurements taken on these entities related by a relevant interaction network. Hence our method is neither organism- nor data-specific. Also, the method is of interest in a wide variety of fields where missing data is a common feature: social sciences, computer vision, remote sensing, speach recognition and of course biological systems. In experiments on synthetic and real biological data, reported in [26], our method demonstrates enhanced results over existing approaches.

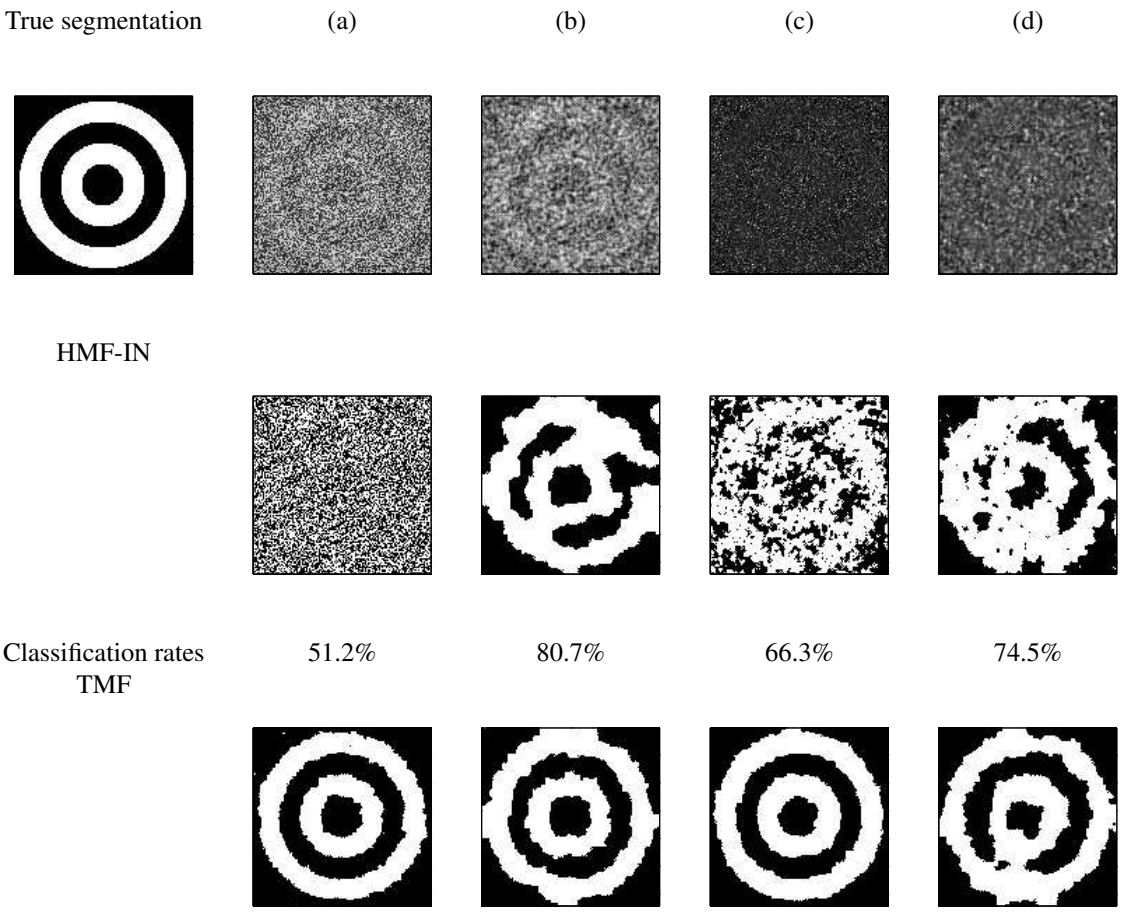


Figure 2. Synthetic image segmentations using a standard Hidden Markov (HMF-IN) model (second row) and our triplet Markov (TMF) model (third row): the true 2-class segmentation is the image in the upper left corner and four different noise models are considered. In (a) class distributions are mixtures of two Gaussians, In (c) observations from class 1 are generated from a Gamma(1,2) distribution and observations from class 2 are obtained by adding 1 to realizations of an Exponential distribution with parameter 1. In (b) and (d) the noisy images are obtained by replacing each pixel value respectively in (a) and (c) by its average with its four nearest neighbors. Classification rates are given below each segmentation results. In the TMF model case, Gaussian components are used to approximate the noise model. The last row gives the number of components K selected using our BIC_{MF} criterion.

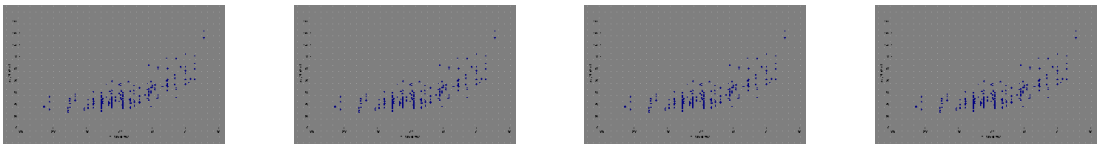
6.2.3. LOCUS: LOcal Cooperative Unified Segmentation of MRI Brain Scans

Participant: Florence Forbes.

Joint work with: Benoit Scherrer, Michel Dojat (Grenoble Institute of Neuroscience) and Christine Garbay (LIG).

MRI brain scan segmentation is a challenging task and has been widely addressed in the last 15 years. Difficulties in automatic segmentation arise from various sources including the size of the data, the low contrast between tissues, the limitations of available prior knowledge, local perturbations such as noise or global perturbations such as intensity nonuniformity. Current approaches share three main characteristics: first, tissue and structure segmentations are considered as two separate tasks whereas they are clearly linked. Second, for a robust to noise segmentation, the Markov Random Field (MRF) probabilistic framework is classically used to introduce spatial dependencies between voxels Third, tissue models are generally estimated globally through the entire volume and do not reflect spatial intensity variations within each tissue, due mainly to biological tissue properties and to MRI hardware imperfections. Only the latter is generally addressed, modeled by the introduction of an explicit so called "bias field" model to estimate. Local segmentation is an attractive alternative. The principle is to compute models in various subvolumes to fit better to local image properties. However, the few local approaches proposed to date are clearly limited: they use local estimation as a preprocessing step only to estimate a bias field model, a training set for statistical local shape modelling , redondant information to ensure consistency and smoothnesss between local estimated models, or an atlas providing a priori local spatial information greedily increasing computational cost. We present in this work [33] an original LOcal Cooperative Unified Segmentation (LOCUS) approach which 1) performs tissue and structure segmentation by distributing a set of cooperating local MRF models through the volume, 2) segments structures by introducing prior localization constraints in a MRF framework and 3) ensures local models consistency and tractable computational time via specific cooperation and coordination mechanisms.

The evaluation was performed using phantoms and real 3T brain scans. It shows good results and in particular robustness to nonuniformity and noise with a low computational cost. Figure 3 shows a visual comparison with two well known approaches, FSL and SPM5, on a very high bias field real 3T brain scan. This image was acquired with a surface coil which provides a high sensitivity in a small region (here the occipital lobe) for functional imaging applications.



(a) (b) (c) (d)

Figure 3. Tissue segmentation of a very high bias field real 3T brain scan (a): segmentations provided by SPM5 (b), FSL (c) and LOCUS (d).

6.2.4. Multimodal MRI segmentation of ischemic stroke lesions

Participant: Florence Forbes.

Joint work with: Benoit Scherrer, Michel Dojat, Yacine Kabir (Grenoble Institute of Neuroscience) and Christine Garbay (LIG).

The problem addressed is the automatic segmentation of stroke lesions on MR multi-sequences. Lesions enhance differently depending on the MR modality and there is an obvious gain in trying to account for various sources of information in a single procedure. To this aim, we propose [32] a multimodal Markov random field

model which includes all MR modalities simultaneously. The results of the multimodal method proposed are compared with those obtained with a mono-dimensional segmentation applied on each MRI sequence separately. We also constructed an Atlas of blood supply territories to help clinicians in the determination of stroke subtypes. Single modality segmentations show as expected that some of the modalities are not or less informative in term of lesion detection and cannot therefore be considered alone. In addition, the modalities information varies with the session. The multimodal approach has the advantage to intrinsically take that into account and to provide satisfactory results in all cases. Further analysis is required. In particular we propose to use the Blood Supply territories Atlas to further assess the performance of the approach.

6.2.5. Joint Markov model for cooperative disparity estimation and object boundary extraction

Participant: Florence Forbes.

Joint work with: Ramya Narasimha, Elise Arnaud, Miles Hansard and Radu Horaud from team Perception, INRIA.

Accurate disparity and object boundary estimation is critical in several applications. In most approaches, these processes are considered as two separate tasks although they are clearly linked: the disparity discontinuities (which are also 3D depth discontinuities) occur usually at object boundaries. However, most disparity estimation algorithms result in disparity discontinuities occurring at improper locations. By “improper" we mean locations which are not at the actual depth discontinuities.

In this work, we build on standard approaches to dense disparity estimation and propose an original approach which simultaneously corrects disparity and finds the object boundaries. These two tasks are dealt with cooperatively, i.e. the presence of disparity discontinuity aids the detection of object boundaries and vice versa. Our approach relies on two assumptions: (i) that the discontinuities in depth are usually at object boundaries (which is true for natural images) (ii) that the disparity discontinuities obtained from naive disparity estimation are usually at the vicinity of actual depth discontinuities. Thus, if we locate the object boundaries which are in the vicinity of the disparity discontinuities – using the gradient map of the image as evidence –, we can correct the disparity values so that they fit closer to the object boundaries. The feedback of boundary estimation on disparity estimation is made through the use of an additional auxiliary field referred to as a *displacement field*. This field suggests the corrections that need to be applied at disparity discontinuities in order that they align with object boundaries, so that disparity discontinuities can then be assumed as representing the object boundaries. The displacement model allows to estimate *directions* in which the discontinuities have to be moved. This information is incorporated in the disparity model so that the disparity values at discontinuities are influenced only by the neighbors in the opposite direction of the displacement. The resulting procedure involves alternation between estimation of disparity and displacement fields in an iterative framework at various scales. When the observation is a set of two stereo images (right and left), we propose a joint probabilistic model of both disparity and displacement fields. Considering the resulting conditional distributions, the formulation reduces to a Markov Random Field (MRF) model on disparities while it reduces to a Markov chain for displacement variables. The disparity-MRF is then optimized using variational mean field and the exact optimization of the Markov chain is carried out using Viterbi algorithm.

The main originality is to define such a model through conditional distributions that can model explicitly relationships between disparity and object boundaries. As a result, we observe a significant gain in disparity and boundary estimations in experiments. The latter show already good results when made with basic image information such as gradient maps. Other monocular cues could be incorporated easily.

As regards, the probabilistic setting itself, we chose to first ignore the parameter estimation issue by fixing them manually. However, a natural future direction of research is to investigate the possibility to incorporate this kind of model in an EM (Expectation Maximization) or variants framework. Besides providing theoretically based parameter estimation, this would also have the advantage to provide a richer framework in which iterative estimation of realizations of the displacement and disparity fields would be replaced by iterative estimation of full distributions for these fields.

6.3. Semi and non-parametric methods

6.3.1. Modelling extremal events

Participants: Stéphane Girard, Laurent Gardes.

Joint work with: 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 dedicated to Weibull-tail distributions (3): kernel estimators [18] and bias correction through exponential regression [14], [15]. 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. The proposed method [16] is based on the comparison between two estimators of quantiles: classical parametric estimators and extreme-value statistics based quantiles.

6.3.2. Conditional extremal events

Participants: Stéphane Girard, Laurent Gardes, Alexandre Lekina.

Joint work with: Cécile Amblard (TimB in TIMC laboratory, Univ. Grenoble 1).

The goal of the PhD thesis of Alexandre Lekina is to contribute to the development of theoretical and algorithmic models to tackle conditional extreme value analysis, *ie* the situation where some covariate information X is recorded simultaneously with a quantity of interest Y. In such a case, 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 will investigate how to combine nonparametric smoothing techniques [39] with extreme-value methods in order to obtain efficient estimators of the conditional tail index and conditional extreme quantiles. 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 (the time). Bivariate examples include the study of extreme rainfalls as a function of the geographical location. Interaction between extreme-value statistics and environmental sciences has been discussed at the Statistical Extremes and Environmental Risk Workshop [29]. The application part of the study will be 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 extreme values. To this aim, a research on some particular copulas [1], [35] has been initiated with Cécile Amblard, since they are the key tool for building multivariate distributions [45].

6.3.3. Boundary estimation

Participants: Stéphane Girard, Laurent Gardes.

Joint work with: Anatoli Iouditski (Univ. Joseph Fourier, Grenoble), Guillaume Bouchard (Xerox, Meylan), Pierre Jacob and Ludovic Menneteau (Univ. Montpellier II) and Alexandre Nazin (IPU, Moscow, Russia).

Two different and complementary approaches are developped.

- **Extreme quantiles approach.** 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 [20], for particular set of points. Our work is 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 [31]. 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 [41], co-directed by Pierre Jacob and Stéphane Girard and in [21] with the consideration of star-shaped supports.

- **Linear programming approach.** 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. 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.

6.3.4. Nuclear plants reliability

Participants: Laurent Gardes, Stéphane Girard.

Joint work with: Nadia Perot, Nicolas Devictor and Michel Marquès (CEA).

One of the main activities of the LCFR (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 (Figure 4).

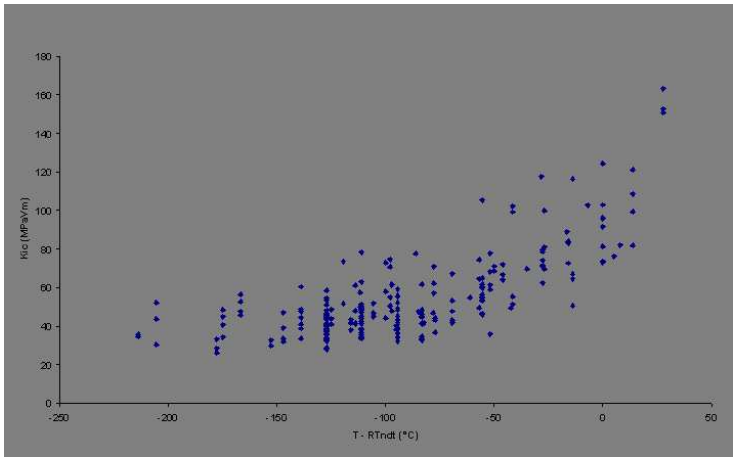


Figure 4. Tenacity as a function of the temperature.

A postdoctoral position on this problem, supported by the CEA, has been opened. Laurent Gardes and Stéphane Girard will co-advise the student. We are currenlty investigating the possibility to sign a research contract on this topic involving MISTIS and the LCFR.

6.3.5. Quantifying uncertainties on extreme rainfall estimations

Participants: Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard.

Joint work with: Gilles Molinié 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-periods than the observed time period (data length). It is then of primary importance to study the sensitivity of the extreme rainfall estimation to the estimation method considered. A preliminary work on this topic is available in [27]. MISTIS got a Ministry grant for a related ANR project (see Section 8.2).

6.3.6. Statistical methods for the analysis of complex remote sensing data

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

Joint work with: Sylvain Douté 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 tera-bits 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 semi-parametric dimension reduction methods such as SIR (Sliced Inverse Regression, [44]) to perform model inversion, in a reasonable computing time, when the number of input observations increases considerably. A preliminary version of this work is presented in [24].

MISTIS got a Ministry grant for a related ANR project (see Section 8.2).

7. Contracts and Grants with Industry

7.1. Contracts

We signed in december 2006 a three-year 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 is co-advised by Stéphane Girard and Jean-Baptiste Durand.

8. Other Grants and Activities

8.1. Regional initiatives

MISTIS participates to 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 Ministry grants for two projects supported by the French National Research Agency (ANR):

- MDCO (Masse de Données et Connaissances) program. This three-year project is called "Visualisation et analyse d'images hyperspectrales multidimensionnelles en Astrophysique" (VAHINEES). It aims at developing physical as well as mathematical models, algorithms, and software able to deal efficiently with hyperspectral multi-angle data but also with any other kind of large hyperspectral dataset (astronomical or experimental). It involves the Observatoire de la Côte d'Azur (Nice), and several universities (Strasbourg I and Grenoble I).
- VMC (Vulnérabilité : Milieux et climats) program. This three-year project is called "Forecast and projection in climate scenario of Mediterranean intense events: Uncertainties and Propagation on environment" (MEDUP) and deals with the quantification and identification of sources of uncertainties associated with the forecast and climate projection for Mediterranean high-impact weather events. The propagation of these uncertainties on the environment is also considered, as well as how they may combine with the intrinsic uncertainties of the vulnerability and risk analysis methods. It involves Météo-France and several universities (Paris VI, Grenoble I and Toulouse III).

MISTIS is also involved into two projects in the Cooperative Research Initiative (ARC) program supported by INRIA:

- The ChromoNet project is coordinated by Marie-France Sagot from team HELIX. It aims at the computational inference and analysis of inter-chromosomal interaction networks. The additional partners are the SSB (Statistiques des Séquences Biologiques) group at INRA and the Nuclear Organisation team at MRC, Imperial College London.
- The SeLMIC project (http://r2-d2.ujf-grenoble.fr/selmic/doku.php) is coordinated by Florence Forbes and aims at developping new statistical methods for the segmentation of multidimensional MR sequences corresponding to different types of MRI modalities and longitudinal data. The applications include the detection of brain abnormalities and more specifically strokes and Multiple Sclerosis lesions. The partners involved are team VisAGeS from INRIA Rennes, the INSERM Unit U594 (Grenoble Institute of Neuroscience) and LIG.

8.3. International initiatives

8.3.1. Europe

J. Blanchet, 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 involved in a European STREP proposal, named POP (Perception On Purpose) coordinated by Radu Horaud from INRIA team Perception. The three-year project started 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 and 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.

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

She is also part of an INRA commitee (CSS MBIA) in charge of evaluating INRA researchers once a year.

F. Forbes and S. Girard are members of the commitees (Commissions de Spécialistes) in charge of examining applications to Faculty member positions respectively at Institut Polytechnique de Grenoble (INPG) and at University Pierre Mendes France (UPMF, Grenoble II) and University Montpellier II.

S. Girard was also involved in the PhD commitee of Céline Vincent from University Montpellier II "Détection de structures tourbillonaires par analyse de données directionnelles" (December 2007).

9.2. University Teaching

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

L. Gardes is faculty member at Univ. P. Mendes-France.

L. Gardes and S. Girard lectured a graduate course on Extreme Value Analysis at Univ. J. Fourier, Grenoble I.

J.B. Durand is faculty member at INPG, Grenoble.

9.3. Conference and workshop committees, invited conferences

Florence Forbes and Gersende Fort were both members of the organizing and scientific committees of the international workshop "New directions in Monte Carlo methods", Fleurance, June 2007.

Stéphane Girard was invited speaker at the workshop "Valeurs extrêmes, méthodes de Monte-Carlo, entropie et information" organized by the GDR Phenix and Isis at ENS Lyon, November 2007.

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] 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.

Year Publications

Articles in refereed journals and book chapters

[9] J. BLANCHET, F. FORBES. *Triplet Markov fields for the classification of complex structure data*, in "IEEE trans. on Pattern Analyis and Machine Intelligence", 2007, To appear.

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

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

[12] G. CELEUX, J.-B. DURAND. *Selecting Hidden Markov Model State Number with Cross-Validated Likelihood*, in "Computational Statistics", to appear, 2007.

[13] C. CHEN, E. DURAND, F. FORBES, O. FRANCOIS. *Bayesian Clustering Algorithms ascertaining spatial population structure: a new computer program and a comparison study*, in "Molecular Ecology Notes", vol. 7, n^{o} 5, 2007, p. 747-756.

[14] J. DIEBOLT, L. GARDES, S. GIRARD, A. GUILLOU. *Bias-reduced estimators of the Weibull tail-coefficient*, in "Test", to appear, 2007.

[15] J. DIEBOLT, L. GARDES, S. GIRARD, A. GUILLOU. *Bias-reduced extreme quantiles estimators of Weibull distributions*, in "Journal of Statistical Planning and Inference", to appear, 2007.

[16] J. DIEBOLT, M. GARRIDO, S. GIRARD. *A Goodness-of-fit Test for the Distribution Tail*, in "Extreme Value Distributions, New-York", M. AHSANULAH, S. KIRMANI (editors), Nova Science, 2007, p. 95–109.

[17] 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", vol. 16, n^{o} 3, 2007, p. 824-837.

[18] L. GARDES, S. GIRARD. *Estimation of the Weibull tail-coefficient with linear combination of upper order statistics*, in "Journal of Statistical Planning and Inference", to appear, 2007.

[19] S. GIRARD, S. IOVLEFF. *Auto-associative models, nonlinear Principal component analysis, manifolds and projection pursuit*, in "Principal Manifolds for Data Visualisation and Dimension Reduction", A. GORBAN, ET AL (editors), vol. 28, LNCSE, Springer-Verlag, 2007, p. 205–222.

[20] S. GIRARD, P. JACOB. *Frontier estimation via kernel regression on high power-transformed data*, in "Journal of Multivariate Analysis", to appear, 2007.

[21] S. GIRARD, L. MENNETEAU. *Smoothed extreme value estimators of non-uniform point processes boundaries with application to star-shaped supports estimation*, in "Communication in Statistics - Theory and Methods", to appear, 2007.

[22] M. VIGNES, F. FORBES. *Gene clustering via integrated Markov models combining individual and pairwise features*, in "IEEE trans. on Computational Biology and Bioinformatics", To appear, 2007.

Publications in Conferences and Workshops

[23] C. BERNARD-MICHEL, C. DE FOUQUET. *Modélisation géostatistique des débits le long des cours d'eau*, in "39èmes Journées de Statistique de la Société Française de Statistique, Angers, France", Avril 2007.

[24] C. BERNARD-MICHEL, S. DOUTÉ, L. GARDES, S. GIRARD. *Estimation of Mars surface physical properties from hyperspectral images using the S.I.R. method*, in "International Symposium on Applied Stochastic Models and Data Analysis, Chania, Crete", mai 2007.

[25] J. BLANCHET, F. FORBES, M. VIGNES. *Clustering with incomplete dependent data*, in "39èmes Journées de Statistique de la Société Française de Statistique, Angers, France", Avril 2007.

[26] J. BLANCHET, M. VIGNES. *Combined expression data with missing values and gene interaction network analysis: a Markovian integrated approach*, in "7th IEEE BIBE Conference, Boston, USA", 2007, p. 366-373.

[27] C. CONTEDUCA, G. MOLINIÉ, S. GIRARD, L. GARDES. *Severe storms in mountainous Mediterranean regions: Uncertainties on extreme rainfall estimations*, in "9th Plinius Conference on Mediterranean Storms, Varenna, Italie", septembre 2007.

[28] J.-B. DURAND. *Calcul de probabilités et estimation dans des modèles de Markov cachés graphiques*, in "39èmes Journées de Statistique de la Société Française de Statistique, Angers, France", Avril 2007.

[29] L. GARDES, S. GIRARD. *Nonparametric estimation of the conditional tail index*, in "Statistical Extremes and Environmental Risk Workshop, Lisbonne", février 2007, p. 47–50.

[30] L. GARDES, S. GIRARD. *Some nonparametric estimators of the conditional tail index*, in "Fifth International Symposium on Extreme Value Analysis, Berne", juillet 2007.

[31] S. GIRARD, P. JACOB. *Boundary estimation via regression on high power-transformed data*, in "56th Session of the International Statistical Institute, Lisbonne", aout 2007.

[32] Y. KABIR, M. DOJAT, B. SCHERRER, F. FORBES, C. GARBAY. *Multimodal MRI segmentation of ischemic stroke lesions*, in "EMBC, Lyon, France", 2007.

[33] B. SCHERRER, M. DOJAT, F. FORBES, C. GARBAY. *LOCUS : LOcal Cooperative Unified Segmentation of MRI Brain Scans*, in "10th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI07, Brisbane, Australia", 2007.

[34] B. SCHERRER, M. DOJAT, F. FORBES, C. GARBAY. *MRF Agent Based Segmentation : Application to MRI Brain Scans*, in "11th Conference on Artificial Intelligence In Medicine, AIME07, Amsterdam, Netherlands", 2007.

References in notes

[35] 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.

[36] 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.

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

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

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

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

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

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

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

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

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

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