



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

*Project-Team MISTIS*

*Modelling and Inference of Complex and  
Structured Stochastic Systems*

*Grenoble - Rhône-Alpes*

THEME COG

*Activity*  
*R* *eport*

2008



## Table of contents

<b>1. Team</b>	<b>1</b>
<b>2. Overall Objectives</b>	<b>1</b>
<b>3. Scientific Foundations</b>	<b>2</b>
3.1. Mixture models	2
3.2. Markov models	2
3.3. Functional Inference, semi and non-parametric methods	3
3.3.1. Modelling extremal events	3
3.3.2. Level sets estimation	4
3.3.3. Dimension reduction	5
<b>4. Application Domains</b>	<b>5</b>
4.1. Image Analysis	5
4.2. Biology, Environment and Medicine	5
4.3. Reliability	5
<b>5. Software</b>	<b>6</b>
5.1. The HDDA and HDDC toolboxes	6
5.2. The Extremes freeware	6
5.3. The SpaCEM <sup>3</sup> program	6
5.4. The FASTRUCT software	7
5.5. The TESS software	7
<b>6. New Results</b>	<b>7</b>
6.1. Mixture models	7
6.1.1. Taking into account the curse of dimensionality.	7
6.1.2. Audio-visual object localization using binaural and binocular cues	8
6.2. Markov models	9
6.2.1. Cooperative clustering	9
6.2.2. Fully Bayesian Joint Model for MR Brain Scan Tissue and Structure Segmentation	9
6.2.3. Brain lesions segmentation from multiple MR sequences	10
6.3. Semi and non-parametric methods	12
6.3.1. Modelling extremal events	12
6.3.2. Conditional extremal events	12
6.3.3. Level sets estimation	12
6.3.4. Dimension reduction	12
6.3.5. Nuclear plants reliability	13
6.3.6. Quantifying uncertainties on extreme rainfall estimations	13
6.3.7. Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression.	14
6.3.8. Statistical analysis of hyperspectral multi-angular data from Mars	15
<b>7. Contracts and Grants with Industry</b>	<b>15</b>
<b>8. Other Grants and Activities</b>	<b>15</b>
8.1. Regional initiatives	15
8.2. National initiatives	16
8.3. International initiatives	16
8.3.1. Europe	16
8.3.2. North Africa	17
8.3.3. North America	17
<b>9. Dissemination</b>	<b>17</b>
9.1. Leadership within scientific community	17
9.2. University Teaching	17
<b>10. Bibliography</b>	<b>17</b>



# 1. Team

## Research Scientist

Florence Forbes [ CR, INRIA ]  
Stéphane Girard [ CR, INRIA, HdR ]  
Gersende Fort [ 20%, Research scientist, CNRS, Paris ]

## Faculty Member

Laurent Gardes [ UPMF, Grenoble ]  
Jean-Baptiste Durand [ 60%, Faculty member, INPG, Grenoble ]  
Marie-José Martinez [ UPMF, Grenoble, since October 2008 ]

## Technical Staff

Sophie Chopart [ Research Engineer ]

## PhD Student

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 ]  
Lamia Azizi [ INRA, co-advised by F. Forbes and S. Girard ]

## Post-Doctoral Fellow

Caroline Bernard-Michel [ INRIA ]  
Senan Doyle [ INRIA ]  
Mathieu Fauvel [ INRIA, since September 2008 ]

## Administrative Assistant

Marie-Anne Dauphin

# 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:** Lamiae Azizi, Senan Doyle, Jean-Baptiste Durand, Florence Forbes, Gersende Fort, Stéphane Girard, Vasil Khalidov.

In a first approach, we consider statistical parametric models,  $\theta$  being the parameter possibly multi-dimensional usually unknown and to be estimated. We consider cases where the data naturally divide into observed data  $y = y_1, \dots, y_n$  and unobserved or missing data  $z = z_1, \dots, z_n$ . 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,

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

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 learning, statistical pattern recognition, stochastic algorithms.*

**Participants:** Lamiae Azizi, Senan Doyle, Jean-Baptiste Durand, Florence Forbes, Gersende Fort, Vasil Khalidov.

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, functional estimation.*

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

We also consider methods which do not assume a parametric model. The approaches are non-parametric in the sense that they do not require the assumption of a prior model on the unknown quantities. This property is important since, for image applications for instance, it is very difficult to introduce sufficiently general parametric models because of the wide variety of image contents. Projection methods are then a way to decompose the unknown quantity on a set of functions (*e.g.* wavelets). Kernel methods which rely on smoothing the data using a set of kernels (usually probability distributions), are other examples. Relationships exist between these methods and learning techniques using Support Vector Machine (SVM) as this appears in the context of *level-sets estimation*, see section 3.3.2. Such non-parametric methods have become the cornerstone when dealing with functional data [46]. 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 (section 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 [51] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [45], which is based on the modelling of distribution tails, see section 3.3.1. It differs from traditional 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  $X_{1,n} \leq \dots \leq X_{n,n}$  denote  $n$  ordered observations from a random variable  $X$  representing some quantity of interest. A  $p_n$ -quantile of  $X$  is the value  $x_{p_n}$  such that the probability that  $X$  is greater than  $x_{p_n}$  is  $p_n$ , *i.e.*  $P(X > x_{p_n}) = p_n$ . When  $p_n < 1/n$ , such a quantile is said to be extreme since it is usually greater than the maximum observation  $X_{n,n}$  (see Figure 1).

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:

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

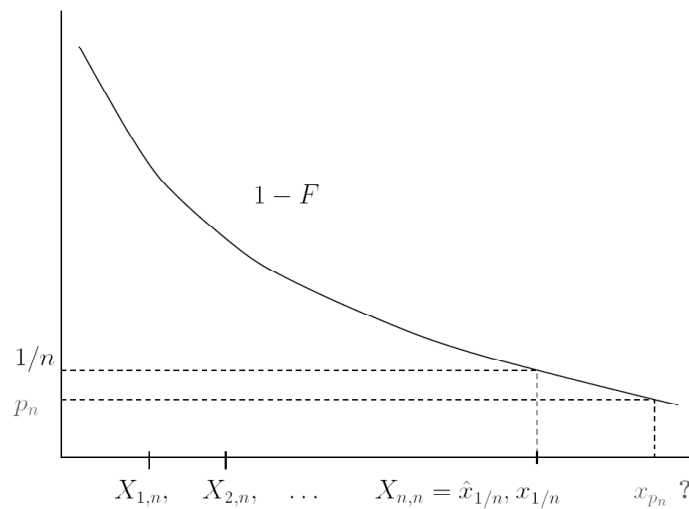


Figure 1. The curve represents the survival function  $x \rightarrow P(X > x)$ . The  $1/n$ -quantile is estimated by the maximum observation so that  $\hat{x}_{1/n} = X_{n,n}$ . As illustrated in the figure, to estimate  $p_n$ -quantiles with  $p_n < 1/n$ , it is necessary to extrapolate beyond the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

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 developed so far. Such models are often referred to as heavy-tail models since the probability of extreme events decreases at a polynomial rate to zero. More generally, the problems that we address are part of the risk management theory. For instance, in reliability, the distributions of interest are included in a semi-parametric family whose tails are decreasing exponentially fast. These so-called Weibull-tail distributions [10] are defined by their survival distribution function:

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

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. We also consider the case where the observations were recorded with a covariate information. In this case, the extreme-value index and the  $p_n$ -quantile are functions of the covariate. We propose estimators of these functions by using a moving window approach.

### 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. Level sets estimation can also be formulated as a linear programming problem [11]. In this context, estimates are sparse since they involve only a small fraction of the dataset, called the set of support vectors.



### 3.3.3. Dimension reduction

Our work on high dimensional data 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 distortions can lead to highly non linear data. Two directions are investigated. First, non-linear PCAs can be proposed, leading to semi-parametric dimension reduction methods [50]. 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 [43]. 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 [51].

## 4. Application Domains

### 4.1. Image Analysis

**Participants:** Caroline Bernard-Michel, Senan Doyle, Mathieu Fauvel, Florence Forbes, Laurent Gardes, Stéphane Girard, Vasil Khalidov.

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 issues about object and class recognition and about the extraction of visual information from large image data bases. In collaboration with team Perception, we also address various issues in computer vision involving Bayesian modelling and probabilistic clustering techniques. Other applications in medical imaging are natural. We work more specifically on MRI data. We also consider other statistical 2D fields coming from other domains such as remote sensing. Also, in the context of the ANR MDCO project, see section 8.2, we work on hyperspectral multi-angle images.

### 4.2. Biology, Environment and Medicine

**Participants:** Lamiae Azizi, Senan Doyle, Florence Forbes, Laurent Gardes, Stéphane Girard, Vasil Khalidov, Alexandre Lekina.

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 and epidemiology are also considered. Finally, in the context of the ANR VMC project, see section 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 [43]. The HDDA and HDDC toolboxes are available for Matlab and are included into the software MixMod [42].

### 5.2. The Extremes freeware

**Participants:** Sophie Chopart, Laurent Gardes, Stéphane Girard.

**Joint work with:** Diebolt, J. (CNRS) and Garrido, M. (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 [49]. 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. Besides, Sophie Chopart has developed a new interface in C++. The software is now independent of Matlab.

### 5.3. The SpaCEM<sup>3</sup> program

**Participants:** Sophie Chopart, Senan Doyle, Florence Forbes.

The SpaCEM<sup>3</sup> (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<sup>3</sup> 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<sup>3</sup> written in C++ is available at <http://spacem3.gforge.inria.fr>. Sophie Chopart started working on the initial version of the software and included a user interface and other improvements. Also we started adding the possibility to deal with mixtures of Poisson distributions in particular in the context of our application to epidemiology.

## 5.4. The FASTRUCT software

**Participant:** Florence Forbes.

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

The FASTRUCT program is dedicated to the modelling and inference of population structure from genetic data. Bayesian model-based clustering programs have gained increased popularity in studies of population structure since the publication of the software STRUCTURE [53]. 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 [44].

## 5.5. The TESS software

**Participant:** Florence Forbes.

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

TESS is a computer program that implements a Bayesian clustering algorithm for spatial population genetics. It is 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 [47]. 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

# 6. New Results

## 6.1. Mixture models

### 6.1.1. Taking into account the curse of dimensionality.

**Participant:** Stéphane Girard.

**Joint work with:** Bouveyron, C (Université Paris 1) and Celeux, G. (Select, INRIA).

In the PhD work of Charles Bouveyron (co-advised by Cordelia Schmid from the INRIA team LEAR) [43], 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 [3]. 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 [2]. In collaboration with Gilles Celeux and Charles Bouveyron we are currently working on the automatic selection of the discrete parameters of the model. Another part of the work of Charles Bouveyron and Stéphane Girard consists in extending this case to the semi-supervised context or to the presence of label noise.

### 6.1.2. Audio-visual object localization using binaural and binocular cues

**Participants:** Florence Forbes, Vasil Khalidov.

**Joint work with:** Arnaud, E., Hansard, M., Horaud, R. and Narasimha, R. from the INRIA team Perception.

This work takes place in the context of the POP European project (see Section 8.3) 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.

**Geometric and probabilistic fusion of spatial visual and auditory cues.** We first explain how we can combine spatial visual and auditory cues in a geometric and probabilistic framework. This is done in order to address the issues of detecting and localizing objects in a scene that are both seen and heard. To do so, we used binaural and binocular sensors for gathering auditory and visual observations. It is shown that the detection and localization problem can be recast as the task of clustering the audio-visual observations into coherent groups. The proposed probabilistic generative model captures the relations between audio and visual observations. This model maps the data into a common audio-visual 3D representation via a pair of mixture models. The statistical method of choice for solving this problem is cluster analysis. We rely on low-level audio and video features which makes our model more general and less dependent on supervised learning techniques, such as face and speech detectors. The input data consists of  $M$  visual observations  $f = \{f_1, \dots, f_m, \dots, f_M\}$ , and  $K$  auditory observations  $g = \{g_1, \dots, g_k, \dots, g_K\}$ . This data is recorded over a time interval  $[t_1, t_2]$ , which is short enough to ensure that the audio-visual (AV) objects responsible for  $f$  and  $g$  are effectively stationary in space. Then we address the estimation of the AV object sites  $S = \{s_1, \dots, s_n, \dots, s_N\}$ , where each  $s_n$  is described by its 3D coordinates  $(x_n, y_n, z_n)^T$ . Note that in general  $N$  is unknown. A visual observation  $f_m$  is a 3D binocular coordinate  $(u_m, v_m, d_m)^T$ , where  $u$  and  $v$  denote the 2D location in the Cyclopean image. The scalar  $d$  denotes the binocular disparity at  $(u, v)^T$ . Hence, Cyclopean coordinates  $(u, v, d)^T$  are associated with each point  $s = (x, y, z)^T$  in the visible scene. We define a function  $F : R^3 \rightarrow R^3$  that maps  $S$  onto  $f$ . An auditory observation  $g_k$  is represented by an auditory disparity, namely the interaural time difference, or ITD. To relate a location to an ITD value we define a function  $G : R^3 \rightarrow R$  that maps  $S$  on  $g$ . Given an observed ITD we can deduce the surface that should contain the source.

We address the problem of AV localization in the framework of unsupervised clustering. The rationale is that observations form groups that correspond to the different AV objects in the scene. So the problem is recast as a clustering task: an assignment of each observation to one of the clusters should be performed as well as the estimation of cluster parameters, which include the  $N$  3D positions  $s_n$  of AV objects. To account for the presence of observations that are not related to any AV object, we introduce an additional background (outlier) class. Because of the different nature of the observations, clustering is performed via two mixture models respectively in the audio (1D) and video (3D) observation spaces, subject to the common parametrization provided by the positions  $s_n$ . The next step is to devise a procedure that finds the best values for the assignments and for the parameters. One possibility is to use a version of the EM algorithm, as it is explained below.

**Development of statistical methods for cross-modal integration.** Given the probabilistic model defined above, we wish to determine the AV objects that generated the visual and auditory observations, that is to derive values of assignment vectors together with the AV object position vectors  $S$  (which are part of our model unknown parameters). Direct maximum likelihood estimation of mixture models is usually difficult, due to the missing assignments. The Expectation Maximization (EM) algorithm is a general and now standard approach to maximization of the likelihood in missing data problems. In our specific context, difficulties arise from the fact that it is necessary to perform simultaneous optimization in two different observation spaces, auditory and visual. It involves solving a system of non-linear equations which does not yield a closed form solution and the traditional EM algorithm cannot be performed. As an alternative, we considered instances of the Generalized EM (GEM) algorithm which is more flexible and provided good results in our experiments. This work has been published in the ICMI'08 conference [36] where more details as well as experiments can be found.

## 6.2. Markov models

### 6.2.1. Cooperative clustering

**Participant:** Florence Forbes.

**Joint work with:** Scherrer, B. and Dojat, M (Grenoble Institute of Neuroscience).

Clustering is a fundamental data analysis step that consists in producing a partitioning of the individuals to account for the groups existing in the observed data. In this paper, we introduce an additional cooperative aspect and propose a framework for more general tasks. We address cases in which the goal is to produce not a single partitioning but two or more partitionings using cooperation between them. Cooperation is expressed by assuming the existence of two sets of missing assignment variables, representing two sets of labels which are not independent but related in the sense that information on one of them is useful to find the other one. We consider non trivial situations in which Markov random field models are used to deal with additional interactions including dependencies between labels within each label sets. We show that our cooperative setting can be formulated in terms of conditional models and propose then to simplify inference into alternating and cooperative estimation procedures based on variants of the Expectation Maximization (EM) algorithm. We illustrate the advantages of our approach by showing its ability to deal successfully with the complex task of segmenting simultaneously and cooperatively tissues and structures from MRI brain scans. In particular this framework is used in the work described in the next section.

### 6.2.2. Fully Bayesian Joint Model for MR Brain Scan Tissue and Structure Segmentation

**Participant:** Florence Forbes.

**Joint work with:** Scherrer, B., Dojat, M. (Grenoble Institute of Neuroscience) and Garbay, C. (LIG).

Difficulties in automatic MR brain scan segmentation arise from various sources. The nonuniformity of image intensity results in spatial intensity variations within each tissue, which is a major obstacle to an accurate automatic tissue segmentation. The automatic segmentation of subcortical structures is a challenging task as well. It cannot be performed based only on intensity distributions and requires the introduction of *a priori* knowledge. Most of the proposed approaches share two main characteristics. First, tissue and subcortical structure segmentations are considered as two successive tasks and treated relatively independently although they are clearly linked: a structure is composed of a specific tissue, and knowledge about structures locations provides valuable information about local intensity distribution for a given tissue. Second, tissue models are estimated globally through the entire volume and then suffer from imperfections at a local level. Alternative local procedures exist but are either used as a preprocessing step or use redundant information to ensure consistency of local models. Recently, we reported good results using an innovative local and cooperative approach [54]. It performs tissue and subcortical structure segmentation by distributing through the volume a set of local Markov Random Field (MRF) models which better reflect local intensity distributions. Local MRF models are used alternatively for tissue and structure segmentations. Although satisfying in practice, these tissue and structure MRF's do not correspond to a valid joint probabilistic model and are not compatible in that sense. As a consequence, important issues such as convergence or other theoretical properties of the resulting local procedure cannot be addressed. In addition, in [54], cooperation mechanisms between local models are somewhat arbitrary and independent of the MRF models themselves. Our contribution [38] is then to propose a fully Bayesian framework in which we define a joint model that links local tissue and structure segmentations but also the model parameters so that both types of cooperations, between tissues and structures and between local models, are deduced from the joint model and optimal in that sense. Our model has the following main features: 1) cooperative segmentation of both tissues and structures is encoded via a joint probabilistic model specified through conditional MRF models which capture the relations between tissues and structures. This model specifications also integrate external *a priori* knowledge in a natural way; 2) intensity nonuniformity is handled by using a specific parametrization of tissue intensity distributions which induces local estimations on subvolumes of the entire volume; 3) global consistency between local estimations is automatically ensured by using a MRF spatial prior for the intensity distributions parameters. Estimation within our framework is defined as a Maximum A Posteriori (MAP) estimation problem and is carried out by adopting an instance of the Expectation Maximization (EM) algorithm. We show that such a setting can adapt well to our conditional models formulation and simplifies into alternating and cooperative estimation procedures for standard Hidden MRF models. The approach is implemented using a multi-agent framework where each agent computes a local MRF model and cooperates with its neighboring agents for model refinement. The evaluation performed using a previously linearly registered atlas of 17 structures show good results. An illustration is given in Figure 2.

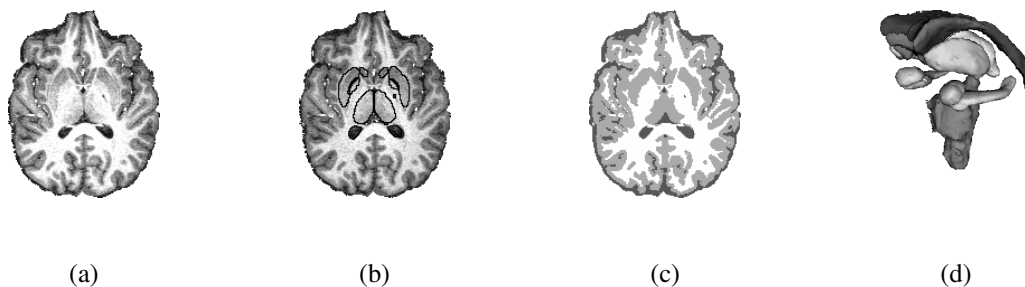


Figure 2. Real 3T brain scan (a). Images (b) and (c): structure segmentation by our method and corresponding improved tissue segmentation. Image (d): 3-D reconstruction of the 17 segmented structures: the two lateral ventricles, caudates, accumbens, putamens, thalamus, pallidums, hippocampus, amygdalas and the brain stem. The computational time was  $< 15min$  after the registration step.

### 6.2.3. Brain lesions segmentation from multiple MR sequences

**Participants:** Florence Forbes, Senan Doyle.



**Joint work with:** Scherrer, B. Dojat, M. (Grenoble Institute of Neuroscience) and Garbay, C. (LIG).

The analysis of MR brain scans is a complex task that is further complicated if the observed data are themselves multi-dimensional as it is the case when several MR channels can provide complementary information and are considered simultaneously. Usually healthy subjects data do not address the same issues as pathological data. This type of data rarely allows the use of automatic or generic approaches. Our goal is to extend our current framework to MRIs with Multiple Sclerosis lesions and stroke lesions. We address the issue of fusing the output of multiple MR sequences to robustly and accurately segment brain lesions. A key capability for radiologists is to delineate lesions out from the rest of the brain tissues. To achieve this goal, radiologists make usually use of multiple MR sequences. The use of multiple sequences not only provides more measurements when segmenting the brain into regions, but crucially, different sequences may be complementary in that one may succeed when another fails. To achieve the same goal automatically and robustly is not an easy task. Overall system performance may be improved in two main ways, either by enhancing the processing of each individual sequence, or by improving the scheme for integrating the information from the different sequences. The contributions of this work concern the latter. We developed a model in which weights can be introduced to account for the relative importance of each modality and propose a variant of the EM algorithm in a Bayesian framework to estimate these weights iteratively and derive a segmentation of the lesions under consideration. Promising results are observed on patients with Multiple Sclerosis lesions (see Figure 3).

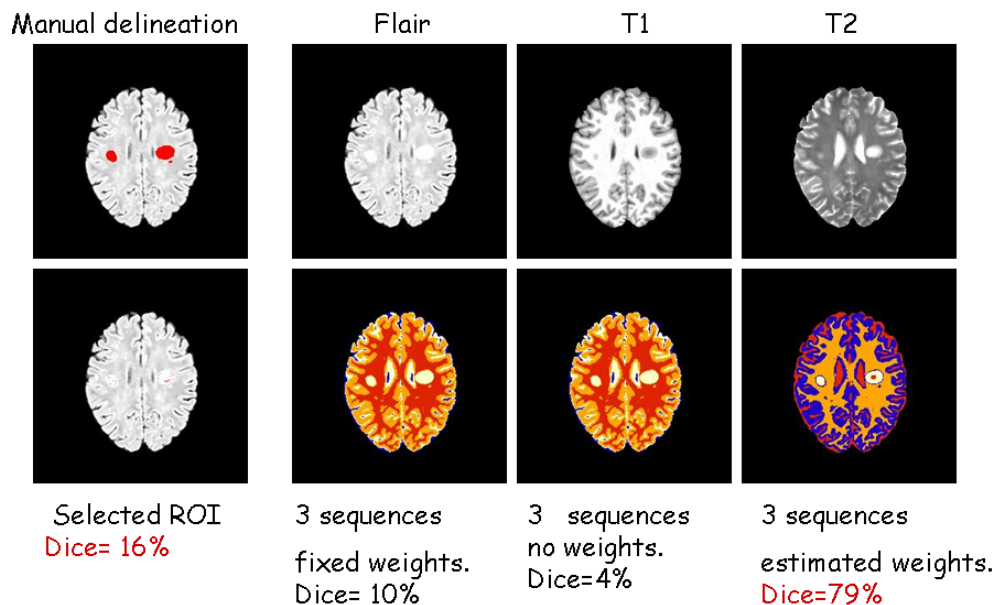


Figure 3. Segmentation of MS lesions from 3 MR images. Our weighted EM procedure (last column) is able to recover a large part of the lesions starting from an initial poor selected ROI (first column). The second and third columns show that weighting is important and must be done adaptively.

## 6.3. Semi and non-parametric methods

### 6.3.1. Modelling extremal events

**Participants:** Stéphane Girard, Laurent Gardes.

**Joint work with:** Guillou, A. (Univ. Strasbourg), and Diebolt, J. (CNRS, Univ. Marne-la-vallée).

Our first achievement is the introduction of a new model of tail distributions depending on a function  $\varphi$  and on an unknown parameter  $\theta$  [40]. This model includes very different distribution tail behaviours from the three classical maximum domains of attraction. In the particular cases of Pareto type tails or Weibull tails, our estimators coincide with classical ones proposed in the literature, thus permitting to retrieve their asymptotic normality in a unified way. Our second achievement is the development of new estimators dedicated to Weibull-tail distributions (3): kernel estimators [19] and bias correction through exponential regression [16], [17].

### 6.3.2. Conditional extremal events

**Participants:** Stéphane Girard, Laurent Gardes, Alexandre Lekina.

**Joint work with:** Amblard, C. (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 combine nonparametric smoothing techniques [46] with extreme-value methods in order to obtain efficient estimators of the conditional tail index [18] and conditional extreme quantiles [41]. 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. 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],[11] has been initiated with Cécile Amblard, since they are the key tool for building multivariate distributions [52].

### 6.3.3. Level sets estimation

**Participants:** Stéphane Girard, Laurent Gardes.

**Joint work with:** Daouia, A. (Univ. Toulouse I), Jacob, P. and Menneteau, L. (Univ. Montpellier II).

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 [21], 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. Besides, we are also working on the extension of our results to more general sets of points. To this end, we focus on the family of conditional heavy tails. An estimator of the conditional tail index has been proposed [18] and the corresponding conditional extreme quantile estimator has been derived [41]. This work has been initiated in the PhD work of Laurent Gardes [48], co-directed by Pierre Jacob and Stéphane Girard and in [22] with the consideration of star-shaped supports.

### 6.3.4. Dimension reduction

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



To overcome the curse of dimensionality arising in high-dimensional regression problems, one way consists in reducing the problem dimension. To this end, Sliced Inverse Regression (SIR) is an interesting solution. The original method, however, requires the inversion of the predictors covariance matrix. In case of collinearity between these predictors or small sample sizes compared to the dimension, the inversion is not possible and a regularization technique has to be used. We thus develop a new approach [13], [12] based on a Fisher Lecture given by R.D. Cook where it is shown that SIR axes can be interpreted as solutions of an inverse regression problem. In this paper, a Gaussian prior distribution is introduced on the unknown parameters of the inverse regression problem in order to regularize their estimation. We show that some existing SIR regularizations can enter our framework, which permits a global understanding of these methods. Three new priors are proposed leading to new regularizations of the SIR method.

This technique has been applied in particular in a collaboration with bioMerieux (see Section 7.1). We co-advised the internship of Lamiae Azizi who applied SIR in the context of quantitation procedures developed at bioMerieux.

### 6.3.5. Nuclear plants reliability

**Participants:** Laurent Gardes, Stéphane Girard.

**Joint work with:** Perot, N., Devictor, N. and Marquès, M. (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).

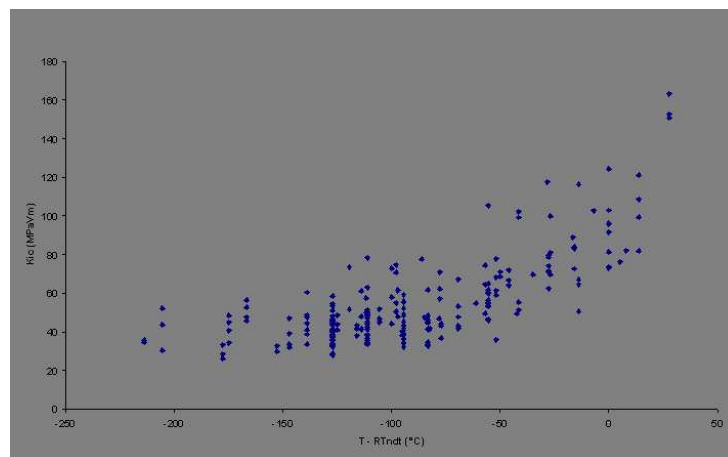


Figure 4. Tenacity as a function of the temperature.

This work is supported by a research contract (from december 2008 to december 2010) involving MISTIS and the LCFR.

### 6.3.6. Quantifying uncertainties on extreme rainfall estimations

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

**Joint work with:** Molinié, G. from Laboratoire d'Etude des Transferts en Hydrologie et Environnement (LTHE), France.

Extreme rainfalls are generally associated with two different precipitation regimes. Extreme cumulated rainfall over 24 hours results from stratiform clouds on which the relief forcing is of primary importance. Extreme rainfall rates are defined as rainfall rates with low probability of occurrence, typically with higher mean return-levels than the maximum observed level. For example Figure 5 presents the return levels for the Cévennes-Vivarais region. 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 [25]. MISTIS got a Ministry grant for a related ANR project (see Section 8.2).

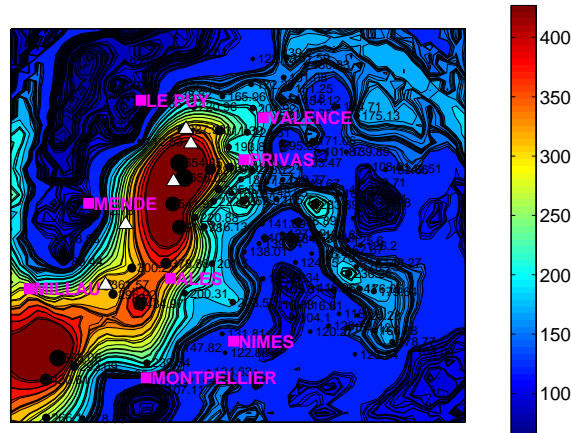


Figure 5. Map of the mean return-levels (in mm) for a period of 10 years.

### 6.3.7. Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression.

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

**Joint work with:** Douté, S. from Laboratoire de Planétologie de Grenoble, France in the context of the VAHINE project (see Section 8.2).

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. Our goal is to estimate the functional relationship  $F$  between some observed spectra and some physical parameters. To this end, a database of synthetic spectra is generated by a physical radiative transfer model and used to estimate  $F$ . The high dimension of spectra is reduced by Gaussian regularized sliced inverse regression (GRSIR) to overcome the curse of dimensionality and consequently the sensitivity of the inversion to noise (ill-conditioned problems). This method is compared with the more classical SVM approach. GRSIR has the advantage of being very fast, interpretable and accurate. Recall that SVM approximates the functional  $F: y = F(x)$  using a solution of the form  $F(x) = \sum_{i=1}^n \alpha_i K(x, x_i) + b$ , where  $x_i$  are samples from the training set,  $K$  a kernel function and  $((\alpha_i)_{i=1}^n, b)$  are the parameters of  $F$  which are estimated during the training process. The kernel  $K$  is used to produce a non-linear function. The SVM training entails

minimization of  $\left[ \frac{1}{n} \sum_{i=1}^{\ell} l(F(x_i), y_i) + \lambda \|F\|^2 \right]$  with respect to  $((\alpha_i)_{i=1}^n, b)$ , and with  $l(F(x), y) = 0$  if  $|F(x) - y| \leq \epsilon$  and  $|F(x) - y| - \epsilon$  otherwise. Prior to running the algorithm, the following parameters need to be fitted:  $\epsilon$  which controls the resolution of the estimation,  $\lambda$  which controls the smoothness of the solution and the kernel parameters ( $\gamma$  for the Gaussian kernel).

### 6.3.8. Statistical analysis of hyperspectral multi-angular data from Mars

**Participants:** Caroline Bernard-Michel, Mathieu Fauvel, Florence Forbes, Laurent Gardes, Stéphane Girard.

**Joint work with:** Douté, S. from Laboratoire de Planétologie de Grenoble, France in the context of the VAHINE project (see Section 8.2).

A new generation of imaging spectrometers is emerging with an additional angular dimension, in addition to the three usual dimensions, two spatial dimensions and one spectral 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. We started 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.

Preliminary experiments carried out by Camille Neels during her 2 month internship in the team pointed out that, previous to any classification task or other analyses, some pre-processing of the images was required. We pointed out the existence in the data of a so-called *spectral smile* issue which we are currently trying to correct. Spectral smile refers to an artefact commonly encountered in spectral images acquired with *Push-broom* spectrometers. It is due to the fact that the wavelength-channel association is not constant across the spatial dimension. Regarding classification tasks, it induces artificial inhomogeneities due to sampling issues.

## 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.

We developed a new collaboration with bioMerieux in Grenoble. We signed a 6 month contract including the co-advising of Lamiae Azizi who was at that time doing an internship at bioMerieux.

We signed a 4 month contract with Veolia-eau in Lyon including the co-advising of Luce Ponsar hired by Veolia for an internship. The goal was to study and possibly detect groups of individuals in time series describing various quantities linked to water consumption and billing in the Lyon area.

## 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" (VAHINE). 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). For more information please visit the associated web site: <http://mistis.inrialpes.fr/vahine/dokuwiki/doku.php>.
- 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). (<http://www.cnrm.meteo.fr/medup/>).

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 developing 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

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.

The final review of the project was held on December 11 and 12, 2008 with in particular a live demo running on the POP audio-visual head regarding multispeaker localisation using binoral and binocular cues. Further details on the project web site <http://perception.inrialpes.fr/POP/>

### 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 committee (CSS MBIA) in charge of evaluating INRA researchers once a year.

S. Girard is member of the committee in charge of examining applications to research scientist (CR) positions at INRIA.

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

F. Forbes was involved in the PhD committee of Benoit Scherrer from INSERM and Grenoble Institut des Neurosciences. The thesis title was "Segmentation des tissus et structures sur les IRM cérébrales: agents markoviens locaux coopératifs et formulation Bayésienne" and the defence held on December 12, 2008.

S. Girard was involved in the PhD committee of Sonia Hedli-Griche from University Grenoble II "Estimation de l'opérateur de régression pour des données fonctionnelles et des erreurs corrélées" (January 2008) and of Matthieu Brucher from University Strasbourg I "Représentations compactes et apprentissage non supervisé de variées non linéaires. Applications au traitement d'image" (October 2008).

### 9.2. University Teaching

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

L. Gardes and M.J. Martinez are faculty members at Univ. P. Mendès-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.

## 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<sup>o</sup> 2, 2005, p. 1–15.

- [2] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional data clustering*, in "Computational Statistics and Data Analysis", vol. 52, 2007, p. 502–519.
- [3] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional discriminant analysis*, in "Communication in Statistics - Theory and Methods", vol. 36, n<sup>o</sup> 14, 2007.
- [4] 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.
- [5] 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<sup>o</sup> 1, 2003, p. 131–144.
- [6] 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.
- [7] F. FORBES, G. FORT. *Combining Monte Carlo and Mean field like methods for inference in hidden Markov Random Fields*, in "IEEE trans. PAMI", vol. 16, n<sup>o</sup> 3, 2007, p. 824–837.
- [8] 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.
- [9] G. FORT, E. MOULINES. *Convergence of the Monte-Carlo EM for curved exponential families*, in "Annals of Statistics", vol. 31, n<sup>o</sup> 4, 2003, p. 1220–1259.
- [10] S. GIRARD. *A Hill type estimate of the Weibull tail-coefficient*, in "Communication in Statistics - Theory and Methods", vol. 33, n<sup>o</sup> 2, 2004, p. 205–234.

## Year Publications

### Articles in International Peer-Reviewed Journal

- [11] C. AMBLARD, S. GIRARD. *A new bivariate extension of FGM copulas*, in "Metrika", To appear, 2008.
- [12] C. BERNARD-MICHEL, L. GARDES, S. GIRARD. *A Note on Sliced Inverse Regression with regularizations*, in "Biometrics", vol. 64, 2008, p. 982–986.
- [13] C. BERNARD-MICHEL, L. GARDES, S. GIRARD. *Gaussian regularized Sliced Inverse Regression*, in "Statistics and Computing", To appear, 2008.
- [14] J. BLANCHET, F. FORBES. *Triplet Markov fields for the supervised classification of complex structure data*, in "IEEE trans. on Pattern Analysis and Machine Intelligence", vol. 30(6), 2008, p. 1055–1067.
- [15] G. CELEUX, J. DURAND. *Selecting Hidden Markov Model State Number with Cross-Validated Likelihood*, in "Computational Statistics", vol. 23(4), 2008, p. 541–564.
- [16] J. DIEBOLT, L. GARDES, S. GIRARD, A. GUILLOU. *Bias-reduced estimators of the Weibull tail-coefficient*, in "Test", vol. 17, 2008, p. 311–331.

- [17] J. DIEBOLT, L. GARDES, S. GIRARD, A. GUILLOU. *Bias-reduced extreme quantiles estimators of Weibull distributions*, in "Journal of Statistical Planning and Inference", vol. 138, 2008, p. 1389–1401.
- [18] L. GARDES, S. GIRARD. *A moving window approach for nonparametric estimation of the conditional tail index*, in "Journal of Multivariate Analysis", vol. 99, 2008, p. 2368–2388.
- [19] L. GARDES, S. GIRARD. *Estimation of the Weibull tail-coefficient with linear combination of upper order statistics*, in "Journal of Statistical Planning and Inference", vol. 139, 2008, p. 1416–1427.
- [20] S. GIRARD, P. JACOB. *A Note on extreme values and kernel estimators of sample boundaries*, in "Statistics and Probability Letters", vol. 78, 2008, p. 1634–1638.
- [21] S. GIRARD, P. JACOB. *Frontier estimation via kernel regression on high power-transformed data*, in "Journal of Multivariate Analysis", vol. 99, 2008, p. 403–420.
- [22] 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", vol. 37, 2008, p. 881–897.
- [23] B. SCHERRER, M. DOJAT, F. FORBES, C. GARBAY. *Agentification of Markov Model Based Segmentation: Application to MRI Brain Scans*, in "Artificial Intelligence in Medicine (AIM)", 2008.
- [24] 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, 2008.

### **International Peer-Reviewed Conference/Proceedings**

- [25] S. ANQUETIN, B. BOUDEVILLAIN, D. CERESSETTI, J. CREUTIN, A. GODART, B. HINGRAY, G. MOLINIÉ, E. LEBLOIS, C. BERNARD-MICHEL, S. GIRARD, L. GARDES. *Rainfall features, forcing and estimation over the Cévennes-Vivarais region*, in "2th HyMeX workshop, Palaiseau, France", juin 2008.
- [26] E. ARNAUD, H. CHRISTENSEN, Y. LU, J. BARKER, V. KHALIDOV, M. HANSARD, B. HOLVECK, H. MATHIEU, R. NARASIMHA, E. TAILLANT, F. FORBES, R. HORAUD. *The CAVA corpus : synchronised stereoscopic and binaural datasets with head movements*, in "ACM/IEEE International Conference on Multimodal Interfaces (ICMI 08)", 2008, p. 109-116.
- [27] J. BENEDIKTSSON, J. CHANUSSOT, M. FAUVEL. *Adaptive pixel neighborhood definition for the classification of hyperspectral images with support vector machines and composite kernel*, in "15th IEEE International Conference on Image Processing, San Diego, Etats-Unis", octobre 2008.
- [28] C. BERNARD-MICHEL, S. DOUTÉ, L. GARDES, S. GIRARD. *Inverting hyperspectral images with Gaussian Regularized Sliced Inverse Regression*, in "16th European Symposium on Artificial Neural Networks, Bruges, Belgique", avril 2008, p. 463–468.
- [29] C. BERNARD-MICHEL, L. GARDES, S. GIRARD. *Regularization methods for Sliced Inverse Regression*, in "8th International Conference on Operations Research, Havana, Cuba", février 2008.
- [30] C. BOUVEYRON, S. GIRARD. *Robust supervised classification with Gaussian mixtures: learning from data with uncertain labels*, in "Compstat, 18th symposium of the IASC, Porto, Portugal", aout 2008.



- [31] V. CIRIZA, L. DONINI, J. DURAND, S. GIRARD. *A statistical model for optimizing power consumption of printers*, in "XIG R & T Conference, Xerox Corporation, Webster, USA", mai 2008.
- [32] V. CIRIZA, L. DONINI, J. DURAND, S. GIRARD. *A statistical model for optimizing power consumption of printers*, in "Joint Meeting of the Statistical Society of Canada and the Société Française de Statistique, Ottawa, Canada", mai 2008.
- [33] L. GARDES, S. GIRARD, A. LEKINA. *A moving window approach for nonparametric estimation of extreme level curves*, in "18th conference of the International Federation of Operational Research Societies, Sandton, Afrique du Sud", juillet 2008.
- [34] S. GIRARD, P. JACOB. *Frontier estimation via regression on high power-transformed data*, in "Joint Meeting of the Statistical Society of Canada and the Société Française de Statistique, Ottawa, Canada", mai 2008.
- [35] V. KHALIDOV, F. FORBES, M. HANSARD, E. ARNAUD, R. HORAUD. *Audio-Visual clustering for 3D speaker localization*, in "5th joint Workshop on Machine Learning and Multimodal Interaction MLMI 2008, Utrecht, The Netherlands", 2008, p. 86-97.
- [36] V. KHALIDOV, F. FORBES, M. HANSARD, E. ARNAUD, R. HORAUD. *Detection and Localization of 3D Audio-Visual Objects Using Unsupervised Clustering*, in "ACM/IEEE International Conference on Multimodal Interfaces (ICMI 08)", 2008, p. 217-224.
- [37] R. NARASIMHA, E. ARNAUD, F. FORBES, R. HORAUD. *Cooperative Disparity and object boundary estimation*, in "15th IEEE Int. Conf. Imag. Proc. ICIP 08, San Diego, USA", 2008, p. 1784–1787.
- [38] B. SCHERRER, F. FORBES, M. DOJAT, C. GARBAY. *Fully Bayesian Joint Model for MR Brain Scan Tissue and Structure Segmentation. Received the Young Investigator Award in Segmentation*, in "MICCAI 2008, New-York, USA", 2008, p. 1066-74.

### Other Publications

- [39] C. BERNARD-MICHEL, S. DOUTÉ, M. FAUVEL, L. GARDES, S. GIRARD. *Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression*, 2008, <http://hal.inria.fr/inria-00276116/fr/>.
- [40] L. GARDES, S. GIRARD, A. GUILLOU. *On the asymptotic normality of extreme-value estimators in the phi-tail distributions model*, 2008, <http://hal.archives-ouvertes.fr/hal-00340661/fr/>.
- [41] L. GARDES, S. GIRARD, A. LEKINA. *Functional nonparametric estimation of conditional extreme quantiles*, 2008, <http://hal.archives-ouvertes.fr/hal-00289996/fr/>.

### References in notes

- [42] C. BIERNACKI, G. CELEUX, G. GOVAERT, F. LANGROGNET. *Model-Based Cluster and Discriminant Analysis with the MIXMOD Software*, in "Computational Statistics and Data Analysis", vol. 51, n<sup>o</sup> 2, 2006, p. 587–600.
- [43] 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>.



- 
- [44] C. CHEN, F. FORBES, O. FRANCOIS. *FASTRUCT: Model-based clustering made faster*, in "Molecular Ecology Notes", vol. 6, 2006, p. 980–983.
- [45] P. EMBRECHTS, C. KLÜPPELBERG, T. MIKOSH. *Modelling Extremal Events*, Applications of Mathematics, vol. 33, Springer-Verlag, 1997.
- [46] F. FERRATY, P. VIEU. *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Springer, 2006.
- [47] O. FRANCOIS, S. ANCELET, G. GUILLOT. *Bayesian clustering using Hidden Markov Random Fields in spatial genetics*, in "Genetics", 2006, p. 805–816.
- [48] L. GARDES. *Estimation d'une fonction quantile extrême*, Ph. D. Thesis, Université Montpellier 2, october 2003.
- [49] 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>.
- [50] 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 Cergy-Pontoise, octobre 1996.
- [51] K. LI. *Sliced inverse regression for dimension reduction*, in "Journal of the American Statistical Association", vol. 86, 1991, p. 316–327.
- [52] R. NELSEN. *An introduction to copulas*, Lecture Notes in Statistics, vol. 139, Springer-Verlag, New-York, 1999.
- [53] J. PRITCHARD, M. STEPHENS, P. DONNELLY. *Inference of Population Structure Using Multilocus Genotype Data*, in "Genetics", vol. 155, 2000, p. 945–959.
- [54] B. SCHERRER, M. DOJAT, F. FORBES, C. GARBAY. *LOCUS: Local Cooperative Unified Segmentation of MRI brain scans*, in "MICCAI 2007, Brisbane, Australia", 2007, p. 219-227.