INRIA

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 : Optimization, Learning and Statistical Methods

# Table of contents

# 1. Team

**Research Scientists**

Florence Forbes [Team Leader, CR,INRIA, HdR]
Stéphane Girard [CR, INRIA, HdR]

**Faculty Members**

Laurent Gardes [UPMF,Grenoble, HdR]
Jean-Baptiste Durand [INPG, Grenoble, in delegation at INRIA Montpellier]
Marie-José Martinez [UPMF, Grenoble]

**PhD Students**

Vasil Khalidov [INRIA, until November 2010, co-advised by F. Forbes and S. Girard]
Alexandre Lekina [INRIA, until October 2010, co-advised by L. Gardes and S. Girard]
Lamiae Azizi [INRA, co-advised by F. Forbes and S. Girard]
Laurent Donini [Cifre contract with Xerox, until May 2010, co-advised by J.B. Durand and S. Girard]
Jonathan El-Methni [INRIA, from October 2010, co-advised by L. Gardes and S. Girard]
El-Hadji Deme [Université Gaston Berger, Sénégal, from October 2010]
Christine Bakhous [INRIA, from November 2010, co-advised by F. Forbes and M. Dojat (GIN)]

**Post-Doctoral Fellows**

Senan James Doyle [INRIA]
Darren Wraith [INRIA]
Julie Carreau [UJF, until May 2010]
Mathieu Fauvel [INRIA, until August 2010]
Kai Qin [INRIA, since April 2010]
Eugen Ursu [INRIA, until August 2010]
Laure Amate [CNRS, from February 2010]
Lotfi Chaari [INRIA, from November 2010]

**Administrative Assistants**

Imma Presseguer [since August 2010]
Patricia Oddos [until July 2010]

**Other**

Eric Frichot [INRIA, Intern from April to June 2010]

# 2. Overall Objectives

## 2.1. Introduction

The MISTIS team aims to develop statistical methods for dealing with complex problems or data. Our applications consist mainly of image processing and spatial data problems with some applications in biology and medicine. Our approach is based on the statement that complexity can be handled by working up from simple local assumptions in a coherent way, defining a structured model, and that is the key to modelling, computation, inference and interpretation. The methods we focus on involve mixture models, Markov models, and, more generally, hidden structure models identified by stochastic algorithms on one hand, and semi and non-parametric methods on the other hand.

Hidden structure models are useful for taking into account heterogeneity in data. They concern many areas of statistical methodology (finite mixture analysis, hidden Markov models, random effect models, etc). Due to their missing data structure, they induce specific difficulties for both estimating the model parameters and assessing performance. The team focuses on research regarding both aspects. We design specific algorithms for estimating the parameters of missing structure models and we propose and study specific criteria for choosing the most relevant missing structure models in several contexts.

Semi- and non-parametric methods are relevant and useful when no appropriate parametric model exists for the data under study either because of data complexity, or because information is missing. The focus is on functions describing curves or surfaces or more generally manifolds rather than real valued parameters. This can be interesting in image processing for instance where it can be difficult to introduce parametric models that are general enough (e.g. for contours).

## 2.2. Highlights

The 17th Working Group on Model-Based Clustering was organized in Grenoble in July 2010. The organizing committee consisted of G. Celeux (INRIA Futurs), F. Forbes (Mistis), B. Murphy (Univ. College Dublin) and A. Raftery (Univ. of Washington). F. Forbes was in charge of the local organization.

# 3. Scientific Foundations

## 3.1. Mixture models

**Participants:** Lamiae Azizi, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Jean-Baptiste Durand, Florence Forbes, Stéphane Girard, Vasil Khalidov, Marie-José Martinez, Darren Wraith.

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 divides into observed data $y = y_1, ..., y_n$ and unobserved or missing data $z = z_1, ..., z_n$. The missing data $z_i$ represents for instance the memberships of one of a set of $K$ alternative categories. The distribution of an observed $y_i$ can be written as a finite mixture of distributions,

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

These models are interesting in that they may point out hidden variable responsible for most of the observed variability and so that the observed variables are *conditionally* independent. Their estimation is often difficult due to the missing data. The Expectation-Maximization (EM) algorithm is a general and now standard approach to maximization of the likelihood in missing data problems. It provides parameter estimation but also values for missing data.

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

## 3.2. Markov models

**Participants:** Lamiae Azizi, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Jean-Baptiste Durand, Florence Forbes, Vasil Khalidov, Darren Wraith.

Graphical modelling provides a diagrammatic representation of the logical structure of a joint probability distribution, in the form of a network or graph depicting the local relations among variables. The graph can have directed or undirected links or edges between the nodes, which represent the individual variables. Associated with the graph are various Markov properties that specify how the graph encodes conditional independence assumptions.

It is the conditional independence assumptions that give graphical models their fundamental modular structure, enabling computation of globally interesting quantities from local specifications. In this way graphical models form an essential basis for our methodologies based on structures.

The graphs can be either directed, e.g. Bayesian Networks, or undirected, e.g. Markov Random Fields. The specificity of Markovian models is that the dependencies between the nodes are limited to the nearest neighbor nodes. The neighborhood definition can vary and be adapted to the problem of interest. When parts of the variables (nodes) are not observed or missing, we refer to these models as Hidden Markov Models (HMM). Hidden Markov chains or hidden Markov fields correspond to cases where the $z_i$'s in (1) are distributed according to a Markov chain or a Markov field. They are a natural extension of mixture models. They are widely used in signal processing (speech recognition, genome sequence analysis) and in image processing (remote sensing, MRI, etc.). Such models are very flexible in practice and can naturally account for the phenomena to be studied.

Hidden Markov models are very useful in modelling spatial dependencies but these dependencies and the possible existence of hidden variables are also responsible for a typically large amount of computation. It follows that the statistical analysis may not be straightforward. Typical issues are related to the neighborhood structure to be chosen when not dictated by the context and the possible high dimensionality of the observations. This also requires a good understanding of the role of each parameter and methods to tune them depending on the goal in mind. Regarding estimation algorithms, they correspond to an energy minimization problem which is NP-hard and usually performed through approximation. We focus on a certain type of methods based on the mean field principle and propose effective algorithms which show good performance in practice and for which we also study theoretical properties. We also propose some tools for model selection. Eventually we investigate ways to extend the standard Hidden Markov Field model to increase its modelling power.

## 3.3. Functional Inference, semi- and non-parametric methods

**Participants:** Julie Carreau, El-Hadji Deme, Jonathan El-Methni, Mathieu Fauvel, Laurent Gardes, Stéphane Girard, Alexandre Lekina, Eugen Ursu.

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 [48]. This is the case, for instance, when observations are curves. They enable us to model the data without a discretization step. More generally, these techniques are of great use for *dimension reduction* purposes (section 3.3.3). They enable reduction of the dimension of the functional or multivariate data without assumptions on the observations distribution. Semi-parametric methods refer to methods that include both parametric and non-parametric aspects. Examples include the Sliced Inverse Regression (SIR) method [55] which combines non-parametric regression techniques with parametric dimension reduction aspects. This is also the case in *extreme value analysis* [47], which is based on the modelling of distribution tails (see section 3.3.1). It differs from traditional statistics which focuses on the central part of distributions, *i.e.* on the most probable events. Extreme value theory shows that distribution tails can be modelled by both a functional part and a real parameter, the extreme value index.

### *3.3.1. Modelling extremal events*

Extreme value theory is a branch of statistics dealing with the extreme deviations from the bulk of probability distributions. More specifically, it focuses on the limiting distributions for the minimum or the maximum of a large collection of random observations from the same arbitrary distribution. Let $X_{1,n} \leq ... \leq X_{n,n}$ denote $n$ ordered observations from a random variable $X$ representing some quantity of interest. A $p_n$-quantile of $X$ is the value $x_{p_n}$ such that the probability that $X$ is greater than $x_{p_n}$ is $p_n$, *i.e.* $P(X > x_{p_n}) = p_n$. When $p_n < 1/n$, such a quantile is said to be extreme since it is usually greater than the maximum observation $X_{n,n}$ (see Figure 1).

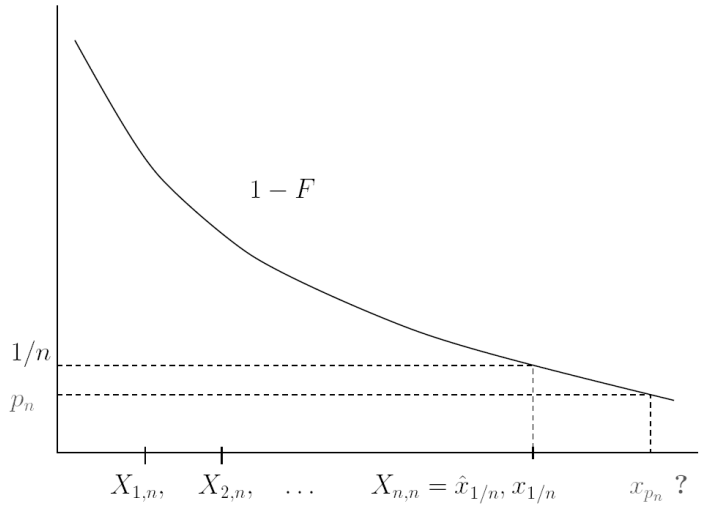


*Figure 1. The curve represents the survival function $x \to P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\widehat{x}_{1/n} = X_{n,n}$. As illustrated in the figure, to estimate $p_n$-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.*

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

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

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

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

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

$$P(X > x) = \exp\{-x^{\theta}\ell(x)\},\ x > x_0 > 0. \tag{4}$$

Gaussian, gamma, exponential and Weibull distributions, among others, are included in this family. An important part of our work consists in establishing links between models (2) and (4) in order to propose new estimation methods. We also consider the case where the observations were recorded with a covariate information. In this case, the extreme-value index and the $p_n$-quantile are functions of the covariate. We propose estimators of these functions by using moving window approaches, nearest neighbor methods, or kernel estimators.

### *3.3.2. Level sets estimation*

Level sets estimation is a recurrent problem in statistics which is linked to outlier detection. In biology, one is interested in estimating reference curves, that is to say curves which bound $90\%$ (for example) of the population. Points outside this bound are considered as outliers compared to the reference population. Level sets estimation can be looked at as a conditional quantile estimation problem which benefits from a non-parametric statistical framework. In particular, boundary estimation, arising in image segmentation as well as in supervised learning, is interpreted as an extreme level set estimation problem. Level sets estimation can also be formulated as a linear programming problem. In this context, estimates are sparse since they involve only a small fraction of the dataset, called the set of support vectors.

### *3.3.3. Dimension reduction*

Our work on high dimensional data requires that we face the curse of dimensionality phenomenon. Indeed, the modelling of high dimensional data requires complex models and thus the estimation of high number of parameters compared to the sample size. In this framework, dimension reduction methods aim at replacing the original variables by a small number of linear combinations with as small as a possible loss of information. Principal Component Analysis (PCA) is the most widely used method to reduce dimension in data. However, standard linear PCA can be quite inefficient on image data where even simple image distorsions can lead to highly non-linear data. Two directions are investigated. First, non-linear PCAs can be proposed, leading to semi-parametric dimension reduction methods [52]. 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 [44]. Such models can then be used in a Mixtures or Markov framework for classification purposes. Another approach consists in combining dimension reduction, regularization techniques, and regression techniques to improve the Sliced Inverse Regression method [55].

# 4. Application Domains

## 4.1. Image Analysis

**Participants:** Christine Bakhous, Lotfi Chaari, Senan James Doyle, Mathieu Fauvel, Florence Forbes, Eric Frichot, Laurent Gardes, Stéphane Girard, Vasil Khalidov, Kai Qin, Darren Wraith.

As regards applications, several areas of image analysis can be covered using the tools developed in the team. More specifically, in collaboration with the Perception team, we address various issues in computer vision involving Bayesian modelling and probabilistic clustering techniques. Other applications in medical imaging are natural. We work more specifically on MRI data, in collaboration with the Grenoble Institute of Neuroscience (GIN) and LNAO from the NeuroSpin center of CEA Saclay (see Sections 5.2 and 6.2.1). We also consider other statistical 2D fields coming from other domains such as remote sensing, in collaboration with Laboratoire de Planétologie de Grenoble. In the context of the ANR MDCO Vahine project, see section 7.2, we work on hyperspectral multi-angle images. In the context of the "pole de competivite" project I-VP, we work of images of PC Boards.

## 4.2. Biology, Environment and Medicine

**Participants:** Lamiae Azizi, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Florence Forbes, Eric Frichot, Laurent Gardes, Stéphane Girard, Vasil Khalidov, Marie-José Martinez, Eugen Ursu, Darren Wraith.

A second domain of applications concerns biology and medecine. We consider the use of missing data models in epidemiology. We also investigated statistical tools for the analysis of bacterial genomes beyond gene detection. Applications in population genetics and neurosiences (Sections 5.2 and 6.2.1) are also considered. Finally, in the context of the ANR VMC project Medup, see section 7.2, we study the uncertainties in the forecasting and climate projection for Mediterranean high-impact weather events.

## 4.3. Reliability

**Participants:** 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 collaboration with Xerox, Meylan.

# 5. Software

## 5.1. The ECMPR software

**Participant:** Florence Forbes.

**Joint work with:** Radu Horaud and Manuel Iguel.

The ECMPR (Expectation Conditional Maximization for Point Registration) package implements [46] [19]. It registers two (2D or 3D) point clouds using an algorithm based on maximum likelihood with hidden variables. The method can register both rigid and articulated shapes. It estimates both the rigid or the kinematic transformation between the two shapes as well as the parameters (covariances) associated with the underlying Gaussian mixture model. It has been registered in APP in 2010 under the GPL license.

## 5.2. The LOCUS software

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

**Joint work with:** Michel Dojat, C. Garbay and B. Scherrer.

The LOCUS software analyses in few minutes a 3D MR brain scan and identifies brain tissues and a large number of brain structures. An image is divided into cubes on each of which a statistical model is applied. This provides a number of local treatments that are then integrated to ensure consistency at a global level. It results a low sensitivity to artefacts. The statistical model is based on a Markovian approach which enables to capture the relations between tissues and structures, to integrate a priori anatomical knowledge and to handle local estimations and spatial correlations. A description and a video of the software are available at the web site http://locus.gforge.inria.fr, which is still under construction. The software is written in C++ (50,000 lines). It has been registered in APP in 2010 under an owner license.

## 5.3. The POPEYE software

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

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

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

## 5.4. The HDDA and HDDC toolboxes

**Participant:** Stéphane Girard.

**Joint work with:** Charles Bouveyron (Université Paris 1) 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 [44]. The HDDA and HDDC toolboxes are available for Matlab and are included into the software MixMod [43]. Recently, a R package has been developped and integrated in The Comprehensive R Archive Network (CRAN). It can be downloaded at the following URL: http://cran.r-project.org/web/packages/HDclassif/.

## 5.5. The Extremes freeware

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

**Joint work with:** Diebolt, J. (CNRS) and Garrido, M. (INRA Clermont-Ferrand-Theix).

The EXTREMES software is a toolbox dedicated to the modelling of extremal events offering extreme quantile estimation procedures and model selection methods. This software results from a collaboration with EDF R&D. It is also a consequence of the PhD thesis work of Myriam Garrido [51]. The software is written in C++ with a Matlab graphical interface. It is now available both on Windows and Linux environments. It can be downloaded at the following URL: http://extremes.gforge.inria.fr/.

## 5.6. The SpaCEM$^3$ program

**Participants:** Lamiae Azizi, Senan James Doyle, Florence Forbes.

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

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

- The unsupervised clustering of dependent objects. Their dependencies are encoded via a graph not necessarily regular and data sets are modelled via Markov random fields and mixture models (eg. MRF and Hidden MRF). Available Markov models include extensions of the Potts model with the possibility to define more general interaction models.
- The supervised clustering of dependent objects when standard Hidden MRF (HMRF) assumptions do not hold (ie. in the case of non-correlated and non-unimodal noise models). The learning and test steps are based on recently introduced Triplet Markov models.

- Selection model criteria (BIC, ICL and their mean-field approximations) that select the "best" HMRF according to the data.
- The possibility of producing simulated data from:
  - general pairwise MRF with singleton and pair potentials (typically Potts models and extensions)
  - standard HMRF, ie. with independent noise model
  - general Triplet Markov models with interaction up to order 2
- A specific setting to account for high-dimensional observations.
- An integrated framework to deal with missing observations, under Missing At Random (MAR) hypothesis, with prior imputation (KNN, mean, etc), online imputation (as a step in the algorithm), or without imputation.

The software is available at http://spacem3.gforge.inria.fr. A user manual in English is available on the web site above together with example data sets.

## 5.7. The FASTRUCT software

**Participant:** Florence Forbes.

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

The FASTRUCT program is dedicated to the modelling and inference of population structure from genetic data. Bayesian model-based clustering programs have gained increased popularity in studies of population structure since the publication of the software STRUCTURE [58]. 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 [45].

## 5.8. The TESS software

**Participant:** Florence Forbes.

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

TESS is a computer program that implements a Bayesian clustering algorithm for spatial population genetics. Is it particularly useful for seeking genetic barriers or genetic discontinuities in continuous populations. The method is based on a hierarchical mixture model where the prior distribution on cluster labels is defined as a Hidden Markov Random Field [49]. Given individual geographical locations, the program seeks population structure from multilocus genotypes without assuming predefined populations. TESS takes input data files in a format compatible to existing non-spatial Bayesian algorithms (e.g. STRUCTURE). It returns graphical displays of cluster membership probabilities and geographical cluster assignments through its Graphical User Interface.

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

- Molecular Ecology Notes 2007

# 6. New Results

## 6.1. Mixture models

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

**Participant:** Stéphane Girard.

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

In the PhD work of Charles Bouveyron (co-advised by Cordelia Schmid from the INRIA LEAR team) [44], we propose new Gaussian models of high dimensional data for classification purposes. We assume that the data live in several groups located in subspaces of lower dimensions. Two different strategies arise:

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

This modelling yields a new supervised classification method called High Dimensional Discriminant Analysis (HDDA) [4]. Some versions of this method have been tested on the supervised classification of objects in images. This approach has been adapted to the unsupervised classification framework, and the related method is named High Dimensional Data Clustering (HDDC) [3].

In collaboration with Gilles Celeux and Charles Bouveyron, we are currently working on the automatic selection of the discrete parameters of the model. The results are submitted for publication [40]. Also, the description of the R package is submitted for publication [39]. An application to the classification of high-dimensional vibrational spectroscopy data has also been developped [20].

### *6.1.2. Information criteria for model selection in the case of multimodal data*

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

**Joint work with:** Radu Horaud from the INRIA Perception team.

A multimodal data setting is a combination of multiple data sets each of them being generated from a different sensors. The data sets live in different physical spaces with different dimensionalities and cannot be embedded in a single common space. We focus on the issue of clustering such multimodal data. This raises the question of how to perform pairwise comparisons between observations living in different spaces. A solution within the framework of Gaussian mixture models and the Expectation-Maximization (EM) algorithm, has been proposed in [21]. Each modality is associated to a modality-specific Gaussian mixture which shares with the others a number of common parameters and a common number of components. Each component corresponds to a common multimodal event that is responsible for a number of observations in each modality. As this number of components is usually unknown, we propose information criteria for selecting this number from the data. We introduce new appropriate criteria based on a penalized maximum likelihood principle. A consistency result for the estimator of the common number of components is given under some assumptions. In practice, the need for a maximum likelihood estimation also requires that we are able to properly initialize the EM algorithm of [21]. We then also propose an efficient initialization procedure. This procedure and the new *conjugate BIC* score we derived are illustrated successfully on a challenging two modality task of detecting and localizing audio-visual objects.

### *6.1.3. Multiple scaled Student distributions with application to clustering*

**Participants:** Florence Forbes, Senan James Doyle, Darren Wraith.

There is an increasingly large literature for statistical approaches to cluster data for a very wide variety of applications. For many applications there has also been an increasing need for approaches to be robust in some sense. For example, in some applications the tails of normal distributions are shorter than appropriate or parameter estimations are affected by atypical observations (outliers). A popular approach proposed for these cases is to fit a mixture of Student distributions (either univariate or multivariate) providing an additional degree of freedom (dof) parameter which can be viewed as a robustness tuning parameter.

An additional advantage of the Student approach is a convenient computational tractability via the use of the EM algorithm with the cluster membership treated as missing variable/data. An additional numerical procedure is then used to find the ML estimate of the degree of freedom.

There are many ways to generalize the Student distribution. Recent approaches such as the skew Student etc.. Much less interest though has focussed on alternative forms for the degree of freedom parameter. The standard student in this regard has one disadvantage: all its marginals are Student but have the same degree of freedom and hence the same amount of tailweight. As noted by Azzalini and Genton in a recent review paper, a simple example is where one variable has Cauchy tails (df=1) and another Gaussian. In this situation, "the single degrees of freedom parameter has to provide a compromise between those two tail behaviours". One solution could be to take a product of independent t-distributions of varying degree of freedom but assuming no correlation between dimensions. For many applications this may however be too strong an assumption. Jones in 2002 proposed a dependent bivariate t distribution with marginals of different degree of freedom but the tractability of the extension to the multivariate case is unclear. Increasingly there has been much research on copula approaches to account for flexible distributional forms but the choice as to which one to use in this case and the applicability to (even) moderate dimensions is not clear.

In this work we propose to extend the Student distribution to allow for the degree of freedom parameter to be estimated differently in each dimension of the parameter space. The key feature of the approach is a decomposition of the covariance matrix which facilitates the separate estimation and also allows for arbitrary correlation between dimensions. The properties of the approach and an assessment of it's performance are outlined on several datasets that are particularly challenging to the standard Student mixture case and also to many alternative clustering approaches.

## 6.2. Markov models

### *6.2.1. Bayesian Weighting of Multiple MR Sequences for Brain Lesion Segmentation*

**Participants:** Florence Forbes, Senan James Doyle, Eric Frichot, Darren Wraith.

**Joint work with:** Michel Dojat (Grenoble Institute of Neuroscience).

A healthy brain is generally segmented into three tissues: cephalo spinal fluid, grey matter and white matter. Statistical based approaches usually aim to model probability distributions of voxel intensities with the idea that such distributions are tissue-dependent. The delineation and quantification of brain lesions is critical to establishing patient prognosis, and for charting the development of pathology over time. Typically, this is performed manually by a medical expert, however automatic methods have been proposed (see [59] for review) to alleviate the tedious, time consuming and subjective nature of manual delineation. Automated or semi-automated brain lesion detection methods can be classified according to their use of multiple sequences, *a priori* knowledge about the structure of normal brain, tissue segmentation models, and whether or not specific lesion types are targeted. A common feature is that most methods are based on the initial identification of *candidate regions* for lesions. In most approaches, normal brain tissue *a priori* maps are used to help identify regions where the damaged brain differs, and the lesion is identified as an outlier. Existing methods frequently make use of complementary information from multiple sequences. For example, lesion voxels may appear atypical in one modality and normal in another. This is well known and implicitly used by neuroradiologists when examining data. Within a mathematical framework, multiple sequences enable the superior estimation of tissue classes in a higher dimensional space.

For multiple MRI volumes, intensity distributions are commonly modelled as multi-dimensional Gaussian distributions. This provides a way to combine the multiple sequences in a single segmentation task but with all the sequences having equal importance. However, given that the information content and discriminative power to detect lesions vary between different MR sequences, the question remains as to how to best combine the multiple channels. Depending on the task at hand, it might be beneficial to weight the various sequences differently.

In this work, rather than trying to detect lesion voxels as outliers from a normal tissue model, we adopt an incorporation strategy whose goal is to identify lesion voxels as an additional fourth component. Such an explicit modelling of the lesions is usually avoided. It is difficult for at least two reasons: 1) most lesions have a widely varying and unhomogeneous appearance (*eg.* tumors or stroke lesions) and 2) lesion sizes can be small (*eg.* multiple sclerosis lesions). In a standard tissue segmentation approach, both reasons usually prevent accurate model parameter estimation resulting in bad lesion delineation. Our approach aims to make this estimation possible by modifying the segmentation model with an additional weight field. We propose to modify the tissue segmentation model so that lesion voxels become inliers for the modified model and can be identified as genuine model components. Compared to *robust estimation* approaches (*eg.* [60]) that consist of down-weighting the effect of outliers on the main model estimation, we aim to increase the weight of candidate lesion voxels to overcome the problem of under-representation of the lesion class.

We introduce weight parameters in the segmentation model and then solve the issue of prescribing values for these weights by developing a Bayesian framework. This has the advantage of avoiding the specification of *ad-hoc* weight values and of enabling the incorporation of expert knowledge through a weight prior distribution. We provide an estimation procedure based on a variational Expectation Maximization (EM) algorithm to produce the corresponding segmentation. Furthermore, in the absence of explicit expert knowledge, we show how the weight prior can be specified to guide the model toward lesion identification. Experiments on artificial and real lesions of various sizes are reported to have demonstrated the good performance of our approach.

These latter experiments have been carried out with a first version of the method that uses diagonal covariance matrices in the Gaussian parts of the model [25], [26]. We extended recently to non-diagonal covariance matrices for a more general formulation. This new formulation is still under validation.

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

**Participants:** Florence Forbes, Lotfi Chaari.

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

The goal is to investigate the possibility of using Variational approximation techniques as an alternative to MCMC-based methods for the joint estimation-detection of brain activity in functional MRI data [56]. We investigated the so-called JDE (Joint Detection Estimation) framework developed by P. Ciuciu and collaborators at NeuroSpin [56] [23], [28] and derived a variational version of it. This new formulation is under validation.

### *6.2.3. Disparity and normal estimation through alternating maximization*

**Participant:** Florence Forbes.

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

In this work [27], we propose an algorithm that recovers binocular disparities in accordance with the surface properties of the scene under consideration. To do so, we estimate the disparity as well as the normals in the disparity space, by setting the two tasks in a unified framework. A novel joint probabilistic model is defined through two random fields to favor both intra-field (within neighboring disparities and neighboring normals) and inter-field (between disparities and normals) consistency. Geometric contextual information is introduced in the models for both normals and disparities. The models are optimized using an appropriate alternating maximization procedure. We illustrate the performance of our approach on synthetic and real data.

### 6.2.4. Consistent detection, localization and tracking of Audio-Visual Objects with Variational EM

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

**Joint work with:** Radu Horaud from the INRIA Perception team.

This work addresses the issue of detecting, locating and tracking objects that are both seen and heard in a scene. We give this problem an interpretation within an unsupervised clustering framework and propose a novel approach based on feature consistency. This model is capable of resolving the observations that are due to detector errors, thus improving the estimation accuracy. We formulate the task as a maximum likelihood estimation problem and perform the inference by a version of the expectation-maximization algorithm, which is formally derived, and which provides cooperative estimates of observation errors, observation assignments, and object tracks. We describe several experiments with single- and multiple- person detection, localization and tracking.

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

**Participants:** Lamiae Azizi, Florence Forbes, Senan James Doyle.

**Joint work with:** David Abrial, Christian Ducrot and Myriam Garrido from INRA Clermont-Ferrand-Theix.

The analysis of the geographical variations of a disease and their representation on a map is an important step in epidemiology. The goal is to identify homogeneous regions in terms of disease risk and to gain better insights into the mechanisms underlying the spread of the disease. Traditionally, the region under study is partitioned into a number of areas on which the observed cases of a given disease are counted and compared to the population size in this area. It has also become clear that spatial dependencies between counts had to be taken into account when analyzing such location-dependent data. One of the most popular approach which has been extensively used in this context, is the so-called BYM model introduced by Besag, York and Mollié in 1991. This model corresponds to a Bayesian hierarchical modelling approach. It is based on an Hidden Markov Random Field (HMRF) model where the latent intrinsic risk field is modelled by a Markov field with continuous state space, namely a Gaussian Conditionally Auto-Regressive (CAR) model. The model inference therefore results in a real-valued estimation of the risk at each location and one of the main reported limitation is that local discontinuities in the risk field are not modelled potentially leading to risk maps that are too smooth. In some cases, coarser representations where areas with similar risk values are grouped are desirable. Grouped representations have the advantage that they provide clearly delimited areas for different risk levels, which is helpful for decision-makers to interpret the risk structure and determine protection measures. Using the BYM model it is possible to derive from the model output such a grouping using, either fixed risk ranges (usually difficult to choose in practice) or a more automated clustering techniques. In any case this post-processing step is likely to be sub-optimal. In this work, we investigate procedures that include such a risk classification.

There have been several attempts to take into account the presence of discontinuities in the spatial structure of the risk. Within hierarchical approaches, one possibility is to move the spatial dependence one level higher in the hierarchy. Green and Richardson in 2002 proposed to replace the continuous risk field by a partition model involving the introduction of a finite number of risk levels and allocations variables to assign each area under study to one of these levels. Spatial dependencies are then taken into account by modelling the allocation variables as a discrete state-space Markov field, namely a spatial Potts model. This results in a discrete HMRF modelling. The general effect is also to recast the disease mapping issue into a clustering task using spatial finite Poisson mixtures. In the same spirit, Fernandez and Green proposed another class of spatial mixture models, in which the spatial dependence is pushed yet one level higher. Of course, the higher the spatial dependencies in the hierarchy the more flexible the model but also the more difficult the parameter estimation. As regards inference, these various attempts have in common the use of simulation intensive Monte Carlo Markov Chain (MCMC) techniques which can present serious difficulties in applying them to large data sets in a reasonable time.

Following the idea of using a discrete HMRF model for disease mapping, we propose to use for inference, as an alternative to simulation-based techniques, an Expectation Maximization framework. This framework is commonly used to solve clustering tasks but leads to intractable computation when considering non-trivial Markov dependencies. However, approximation techniques are available and, among them we propose to investigate variational approximations for their computational efficiency and good performance in practice. In particular, we consider the so-called mean field principle [6] that provides a deterministic way to deal with intractable MRF models and has proven to perform well in a number of applications.

Human disease data usually has this particularity that the populations under consideration are large and the risk values relatively high, say between 0.5 and 1.5. This is not fully representative of epidemiological studies, especially studies of non-contagious diseases in animals. When considering animal epidemiology, we may have to face instead low size populations and risk levels much smaller than 1, typically $10^{-5}$ to $10^{-3}$. Difficulties in applying techniques that work in the first (human) case to data sets in the second (animal) case have not been investigated. In addition, no particular difficulties regarding initialization and model selection are usually reported. This is far from being the case in all practical problems. In this work we propose to go further and to address a number of related issues. More specifically, we investigate the model behavior in more detail. We pay special attention to the main two inherent issues when using EM procedures, namely algorithm initialization and model selection. The EM solution can highly depend on its starting position. We show that simple initializations do not always work, especially for rare disease for which the risks are small. We then propose and compare different initialization strategies in order to get a robust way of initializing for most situations arising in practice.

In addition we build on the standard hidden Markov field model by considering a more general formulation that is able to encode more complex interactions than the standard Potts model. In particular we are able to encode the fact that risk levels in neighboring regions cannot be too different while the standard Potts model penalizes the same way different neighboring risks whatever the amplitude of their difference.

### *6.2.6. Optimization of the consumption of printers using Markov decision processes*

**Participants:** Laurent Donini, Jean-Baptiste Durand, Stéphane Girard.

**Joint work with:** Ciriza, V. and Bouchard, G. (Xerox XRCE, Meylan).

In the context of the PhD thesis of Laurent Donini, we have proposed several approaches to optimize the resources consumed by printers. The first aim of this work is to determine an optimal value of the timeout of an isolated printer, so as to minimize its electrical consumption. This optimal timeout is obtained by modeling the stochastic process of the print requests, by computing the expected consumption under this model, according to the characteristics of the printers, and then by minimizing this expectation with respect to the timeout. Two models are considered for the request process: a renewal process, and a hidden Markov chain. Explicit values of the optimal timeout are provided when possible. In other cases, we provide some simple equation satisfied by the optimal timeout. It is also shown that a model based on a renewal process offers as good results as an empirical minimization of the consumption based on exhaustive search of the timeout, for a largely lower computational cost. This work has been extended to take into account the users' discomfort resulting from numerous shutdowns of the printers, which yield increased waiting time. This has also been extended to printers with several states of sleep, or with separate reservoirs of solid ink. The results are submitted for publication [41].

As a second step, the case of a network of printers has been considered. The aim is to decide on which printer some print request must be processed, so as to minimize the total power consumption of the network of printers, taking into account user discomfort. Our approach is based on Markov Decision Processes (MDPs), and explicit solutions for the optimal decision are not available anymore. Furthermore, to simplify the problem, the timeout values are considered are fixed. The state space is continuous, and its dimension increases linearly with the number of printers, which quickly turns the usual algorithms (*i.e.* value or policy iteration) intractable. This is why different variants have been considered, among which the Sarsa algorithm.

## 6.3. Semi and non-parametric methods

### 6.3.1. Modelling extremal events

**Participants:** Stéphane Girard, Laurent Gardes, Jonathan El-methni, El-Hadji Deme.

**Joint work with:** Guillou, A. (Univ. Strasbourg).

We introduced a new model of tail distributions depending on two parameters $\tau \in [0,1]$ and $\theta > 0$ [17]. This model includes very different distribution tail behaviors from Fréchet and Gumbel maximum domains of attraction. In the particular cases of Pareto type tails ($\tau = 1$) or Weibull tails ($\tau = 0$), our estimators coincide with classical ones proposed in the literature, thus permitting us to retrieve their asymptotic normality in an unified way. Our current work consists in defining an estimator of the parameter $\tau$. This would permit the construction of new estimators of extreme quantiles and to propose a test procedure in order to discriminate between Pareto and Weibull tails.
We are also working on the estimation of the second order parameter $\rho$ (see paragraph 3.3.1). Our goal is to propose a new family of estimators encompassing the existing ones (see for instance [54], [53]). This work is in collaboration with El-Hadji Deme, a PhD student from the Université de Saint-Louis (Sénégal). El-Hadji Deme obtained a one-year mobility grant to work within the Mistis team on extreme-value statistics.

### 6.3.2. Conditional extremal events

**Participants:** Stéphane Girard, Laurent Gardes, Julie Carreau, Alexandre Lekina, Eugen Ursu.

**Joint work with:** Amblard, C. (TimB in TIMC laboratory, Univ. Grenoble I) and Daouia, A. (Univ. Toulouse I)

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 [48] with extreme-value methods in order to obtain efficient estimators of the conditional tail index and conditional extreme quantiles. When the covariate is deterministic (fixed design), moving window and nearest neighbours methods are adopted [18]. When the covariate is random (random design), we focus on kernel methods [15]. Conditional extremes are studied in climatology where one is interested in how climate change over years might affect extreme temperatures or rainfalls. In this case, the covariate is univariate (time). Bivariate examples include the study of extreme rainfalls as a function of the geographical location. The application part of the study is joint work with the LTHE (Laboratoire d'étude des Transferts en Hydrologie et Environnement) located in Grenoble [16].

More future work will include the study of multivariate and spatial extreme values. With this aim, a research on some particular copulas [1] has been initiated with Cécile Amblard, since they are the key tool for building multivariate distributions [57]. The PhD thesis of Jonathan El-methni should address this problem too.

### 6.3.3. Level sets estimation

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

**Joint work with:** Guillou, A. (Univ. Strasbourg), Stupfler, G. (Univ. Strasbourg), P. Jacob (Univ. Montpellier II) and Daouia, A. (Univ. Toulouse I).

The boundary bounding the set of points is viewed as the larger level set of the points distribution. This is then an extreme quantile curve estimation problem. We proposed estimators based on projection as well as on kernel regression methods applied on the extreme values set, for particular set of points [10].
In collaboration with A. Daouia, we investigate the application of such methods in econometrics [37]: A new characterization of partial boundaries of a free disposal multivariate support is introduced by making use of large quantiles of a simple transformation of the underlying multivariate distribution. Pointwise empirical and smoothed estimators of the full and partial support curves are built as extreme sample and smoothed quantiles. The extreme-value theory holds then automatically for the empirical frontiers and we show that some fundamental properties of extreme order statistics carry over to Nadaraya's estimates of upper quantile-based frontiers.

In the PhD thesis of Gilles Stupfler (co-directed by Armelle Guillou and Stéphane Girard), new estimators of the boundary are introduced. The regression is performed on the whole set of points, the selection of the "highest" points being automatically performed by the introduction of high order moments. The results are submitted for publication [42].
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, and the corresponding conditional extreme quantile estimator has been derived [18] in a fixed design setting. The extension to the random design framework is published in [15]. This work has been initiated in the PhD work of Laurent Gardes [50], co-directed by Pierre Jacob and Stéphane Girard.

### *6.3.4. 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 steel tenacity in nuclear plant 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 behavior of the model in the distribution tail. In particular, we are mainly interested in studying the lowest tenacities when the temperature varies (Figure 2).

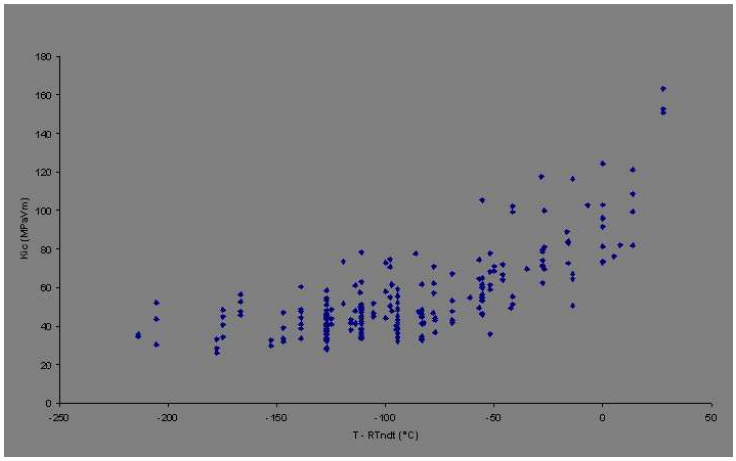


*Figure 2. Tenacity as a function of the temperature.*

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

### *6.3.5. Quantifying uncertainties on extreme rainfall estimations*

**Participants:** Julie Carreau, Eugen Ursu, 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 3 presents the return levels for the Cévennes-Vivarais region obtained in [16]. 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 has been presented in two international workshops on climate [32], [33]. MISTIS got a Ministry grant for a related ANR project (see Section 7.2).

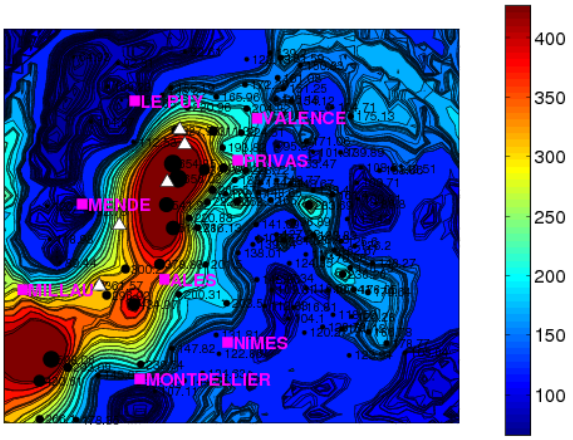


*Figure 3. Map of the mean return-levels (in mm) for a period of 10 years.*

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

**Participants:** 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 7.2).

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 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 $\left[\frac{1}{n}\sum_{i=1}^{\ell} l\left(F(x_i), y_i\right) + \lambda \|F\|^2\right]$ with respect to $((\alpha_i)_{i=1}^{n}, \, b)$, and with $l\left(F(x), y\right) = 0$ 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).

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

**Participants:** 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 7.2).

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 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, and 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. To investigate the use of statistical techniques to deal with these generic sources of complexity, we made preliminary experiments using our HDDC technique on a first set of realistic synthetic 4D spectral data provided by our collaborators from LPG. However, it appeared that this data set was not relevant for our study due to the fact that the simulated angular information provided was not discriminant and could not enable us to draw useful conclusions. Further experiments on other data sets are then necessary.

# 7. Other Grants and Activities

## 7.1. National Actions

MISTIS participates in the weekly statistical seminar of Grenoble. F. Forbes is one of the organizers and several lecturers have been invited in this context.

## 7.2. National initiatives

MISTIS got, for the period 2008-2010, 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 two 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 forecasting 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 three universities (Paris VI, Grenoble I and Toulouse III). (http://www.cnrm.meteo.fr/medup/).

MISTIS is also a partner in a new three-year MINALOGIC project (I-VP for Intuitive Vision Programming) supported by the French Government. The project is led by VI Technology (http://www.vitechnology.com), a world leader in Automated Optical Inspection (AOI) of a broad range of electronic components. The other partners involved are the CMM (Centre de Morphologie Mathematiques) in Fontainebleau, and Pige Electronique in Bourg-Les-Valence. The NOESIS company, which is a leader in the field of image processing and analysis software, in Crolles, is also involved to provide help with software development. The overall goal is to exploit statistical and image processing techniques more intensively to improve defect detection capability and programming time based on existing AOI principles so as to eventually reach a reliable defect detection with virtually zero programming skills and efforts.

MISTIS is also involved in another three-year MINALOGIC project, called OPTYMIST-II, through the co-advising, with Dominique Morche from LETI, of Julie Carreau's post-doctoral subject. The goal is to address variability issues when designing electronic components.

## 7.3. International initiatives

### *7.3.1. North Africa*

S. Girard has joint work with M. El Aroui (ISG Tunis) and El-Hadji Deme (PhD student from the Université de Saint-Louis, Sénégal)

### *7.3.2. North America*

F. Forbes has joint work with C. Fraley and A. Raftery (Univ. of Washington, USA).

### *7.3.3. Europe*

**European STREP HUMAVIPS (2010-13).** MISTIS is involved in a new three-year European project (STREP) started in February 2010. The project is named HUMAVIPS (Humanoids ables with auditory and visual abilities in populated spaces) and was in 2009/10 the only INRIA coordinated project granted in the highly competitive FP7-ICT program of the European Union. The partners involved are the Perception and Mistis teams from INRIA Rhone-alpes (coord.), the Czech Technical University CTU Czech Republic, Aldebaran Robotics ALD France, Idiap Research Institute Switzerland and Bielefeld University BIU Germany. The goal is to develop humanoid robots with integrated audio-visual perception systems and social skills, capable of handling multi-party conversations and interactions with people in realtime. The MISTIS contribution will consist in developing statistical machine learning techniques for interactive robotic applications.

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

M.J. Martinez has joint work with Prof. J. Hinde and E. Holian (National University of Ireland, Galway, Ireland).

# 8. Dissemination

## 8.1. Leadership within scientific community

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

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

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 a member of the committee (Comité de Sélection) in charge of examining applications to Faculty member positions at University Pierre Mendes France (UPMF, Grenoble II).

F. Forbes and S. Girard were elected as members of the bureau of the “Analyse d’images, quantification, et statistique” group in the Société Française de Statistique (SFdS).

S. Girard was selected as an expert for the national fund for the scientific development of Chili (FONDECYT).

S. Girard was selected as an expert by the Research concil of the University of Leuven to evaluate research proposals.

S. Girard was involved in the PhD committees of Dmitri Novikov (Université Montpellier II) and Thi Mong Ngoc Nguyen (Université de Bordeaux).

F. Forbes was involved in the PhD committes of Tomas Crivelli from team VISTA INRIA Rennes, Univ. Rennes I. PhD title: Mixed state Markov models for image motion analysis (March 2010) and of Lotfi Châari from University Paris-Est. PhD subject: Reconstruction d’images médicales d’IRM à l’aide de représentations en ondelettes (November 2010).

F. Forbes was also involved in the HDR committee of Nicolas Wicker, assistant professor at Strasbourg University (December 2010).

## 8.2. Teaching

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

L. Gardes and M.-J. Martinez are faculty members at Univ. Pierre Mendès France, Grenoble II.

L. Gardes and S. Girard lectured a graduate course on extreme value analysis at Univ. Joseph Fourier, Grenoble I.

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

# 9. Bibliography

## Major publications by the team in recent years

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

[2] J. BLANCHET, F. FORBES. *Triplet Markov fields for the supervised classification of complex structure data*, in "IEEE trans. on Pattern Analyis and Machine Intelligence", 2008, vol. 30(6), p. 1055–1067.

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

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

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

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

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

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

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

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

# Publications of the year

## Doctoral Dissertations and Habilitation Theses

[11] F. FORBES. *Models and Inference for structured stochastic systems*, Université Joseph-Fourier - Grenoble I, December 2010, Habilitation à Diriger des Recherches.

[12] L. GARDES. *Contributions la théorie des valeurs extrêmes et la réduction de dimension pour la régression*, Université Joseph-Fourier - Grenoble I, November 2010, Habilitation à Diriger des Recherches, http://hal.inria.fr/tel-00540747/en.

[13] V. KHALIDOV. *Conjugate mixture models for the modelling of visual and auditory perception*, Université Joseph-Fourier - Grenoble I, October 2010.

[14] A. LEKINA. *Estimation non paramétrique des quantiles extrêmes conditionnels*, Université Joseph-Fourier - Grenoble I, October 2010, http://hal.inria.fr/tel-00529476/en.

## Articles in International Peer-Reviewed Journal

[15] A. DAOUIA, L. GARDES, S. GIRARD, A. LEKINA. *Kernel estimators of extreme level curves*, in "Test", 2010, to appear.

[16] L. GARDES, S. GIRARD. *Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels*, in "Extremes", 2010, vol. 13, n° 2, p. 177–204.

[17] L. GARDES, S. GIRARD, A. GUILLOU. *Weibull tail-distributions revisited: a new look at some tail estimators*, in "Journal of Statistical Planning and Inference", 2010, vol. 141, n° 1, p. 429–444.

[18] L. GARDES, S. GIRARD, A. LEKINA. *Functional nonparametric estimation of conditional extreme quantiles*, in "Journal of Multivariate Analysis", 2010, vol. 101, p. 419-433, http://hal.inria.fr/hal-00289996/en.

[19] R. HORAUD, F. FORBES, M. YGUEL, G. DEWAELE, J. ZHANG. *Rigid and Articulated Point Registration with Expectation Conditional Maximization*, in "IEEE Trans. on Pattern Analysis and Machine Intelligence", 2010, To appear.

[20] J. JACQUES, C. BOUVEYRON, S. GIRARD, O. DEVOS, L. DUPONCHEL, C. RUCKEBUSCH. *Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data*, in "Journal of Chemometrics", 2010, to appear.

[21] V. KHALIDOV, F. FORBES, R. HORAUD. *Conjugate Mixture Models for Clustering Multimodal Data*, in "Neural Computation", 2011, vol. 23, n$^{o}$ 2, p. 517-557.

[22] M.-J. MARTINEZ, B. DURAND, D. CALAVAS, C. DUCROT. *Methodological approach for substantiating disease freedom in a small heterogeneous population. Application to ovine scrapie, a disease with a strong genetic susceptibility*, in "Preventive Veterinary Medicine", 2010, vol. 95, p. 108–114.

[23] L. RISSER, T. VINCENT, F. FORBES, J. IDIER, P. CIUCIU. *Min-max extrapolation scheme for fast estimation of 3D Potts field partition functions. Application to the joint detection-estimation of brain activity in fMRI*, in "Special issue of Journal of Signal Processing Systems", 2010.

### International Peer-Reviewed Conference/Proceedings

[24] C. BOUVEYRON, G. CELEUX, S. GIRARD. *Intrinsic Dimension Estimation by Maximum Likelihood in Probabilistic PCA*, in "73rd Annual Meeting of the Institute of Mathematical Statistics", Gothenburg, Sweden, 2010.

[25] F. FORBES, S. DOYLE, D. GARCIA-LORENZO, C. BARILLOT, M. DOJAT. *A Weighted Multi-Sequence Markov Model For Brain Lesion Segmentation*, in "13th International Conference on Artificial Intelligence and Statistics (AISTATS10)", Sardinia, Italy, 13-15 May 2010.

[26] F. FORBES, S. DOYLE, D. GARCIA-LORENZO, C. BARILLOT, M. DOJAT. *Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation*, in "IEEE International Symposium on Biomedical Imaging (ISBI)", Rotterdam, The Netherlands, 14-17 April 2010.

[27] R. NARASIMHA, E. ARNAUD, F. FORBES, R. HORAUD. *Disparity and normal estimation through alternating maximization*, in "international conference on image processing (ICIP)", Hong-Kong Honk Kong, 2010, http://hal.inria.fr/inria-00517864/en.

[28] L. RISSER, T. VINCENT, F. FORBES, J. IDIER, P. CIUCIU. *How to deal with brain deactivations in the joint detection-estimation framework?*, in "Human Brain Mapping (HBM) meeting", Barcelone, Spain, 2010.

### National Peer-Reviewed Conference/Proceedings

[29] D. ABRIAL, L. AZIZI, M. CHARRAS-GARRIDO, F. FORBES. *Approche variationnelle pour la cartographie spatio-temporelle du risque en épidémiologie l'aide de champs de Markov cachés*, in "42èmes Journées de Statistique", France Marseille, France, 2010.

[30] A. DAOUIA, L. GARDES, S. GIRARD, A. LEKINA. *Estimation de courbes de niveaux extrêmes pour des lois à queues lourdes*, in "42èmes Journées de Statistique", France Marseille, France, 2010, http://hal.inria.fr/inria-00494684/en.

[31] M. VIGNES, J. BLANCHET, D. LEROUX, F. FORBES. *Clustering of incomplete, high dimensional and dependent biological data with SpaCEM3*, in "Journee satellite MODGRAPH 2010 de JOBIM", Montpellier, France, 2010.

### Workshops without Proceedings

[32] J. CARREAU, S. GIRARD, E. URSU. *Spatial kernel interpolation for annual rainfall maxima*, in "NICDS Workshop on Statistical Methods for Geographic and Spatial Data in the Management of Natural Resources", Montréal, Canada, mars 2010.

[33] J. CARREAU, S. GIRARD, E. URSU. *Spatial kernel interpolation for annual rainfall maxima*, in "Workshop on metrics and methodologies of estimation of extreme climate events", UNESCO headquarters, Paris, septembre 2010.

[34] S. GIRARD. *On the regularization of the Sliced Inverse Regression*, in "Workshop on Challenging problems in Statistical Learning", Paris, janvier 2010.

[35] J. HINDE, S. DE FREITAS, M.-J. MARTINEZ, C. DEMETRIO, G. PAPAGEORGIOU. *Random effects in cumulative mortality models*, in "XXVth International Biometric Conference", Florianópolis, Brazil, December 2010.

[36] J. HINDE, S. DE FREITAS, M.-J. MARTINEZ, C. DEMETRIO, G. PAPAGEORGIOU. *Random effects in cumulative mortality models*, in "The 2010 Conference of Applied Statistics in Ireland", Portrush, Northern Ireland, May 2010.

### Scientific Books (or Scientific Book chapters)

[37] A. DAOUIA, L. GARDES, S. GIRARD. *Nadaraya's estimates for large quantiles and free disposal support curves*, in "Exploring research frontiers in contemporary statistics and econometrics - Festschrift in honor of L. Simar", I. V. KEILEGOM, P. WILSON (editors), Springer, 2010, to appear.

[38] B. SCHERRER, F. FORBES, C. GARBAY, M. DOJAT. *A joint Bayesian framework for MR brain scan tissue and structure segmentation based on distributed Markovian agents*, in "Computational Intelligence in Healthcare", I. BICHINDARITZ, L. JAIN (editors), Springer-Verlag, Berlin, 2010, To appear.

### Other Publications

[39] L. BERGÉ, C. BOUVEYRON, S. GIRARD. *HDclassif: an R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data*, 2010, http://hal.inria.fr/hal-00541203/en.

[40] C. BOUVEYRON, G. CELEUX, S. GIRARD. *Intrinsic Dimension Estimation by Maximum Likelihood in Probabilistic PCA*, 2010, http://hal.inria.fr/hal-00440372/en.

[41] V. CIRIZA, L. DONINI, J.-B. DURAND, S. GIRARD. *Optimal timeouts for power management under renewal or hidden Markov processes for requests*, 2010, http://hal.archives-ouvertes.fr/hal-00412509/en/.

[42] S. GIRARD, A. GUILLOU, G. STUPFLER. *Frontier estimation with kernel regression on high order moments*, 2010, http://hal.inria.fr/hal-00499369/en.

## References in notes

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

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

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

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

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

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

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

[50] L. GARDES. *Estimation d'une fonction quantile extrême*, Université Montpellier 2, october 2003.

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

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

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

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

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

[56] S. MAKNI, J. IDIER, T. VINCENT, B. THIRION, G. DEHAENE-LAMBERTZ, P. CIUCIU. *A fully Bayesian approach to the parcel-based detection-estimation of brain activity in fMRI.*, in "NeuroImage", 07 2008, vol. 41, n$^{o}$ 3, p. 941-69 [*DOI :* 10.1016/J.NEUROIMAGE.2008.02.017], http://hal-cea.archives-ouvertes.fr/cea-00333624/en/.

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

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

[59] M. SEGHIER, A. RAMLACKHANSINGH, J. CRINION, A. LEFF, C. J. PRICE. *Lesion identification using unified segmentation-normalisation models and fuzzy clustering*, in "Neuroimage", 2008, vol. 41, p. 1253-1266.

[60] K. VAN LEEMPUT, F. MAES, D. VANDERMEULEN, A. COLCHESTER, P. SUETENS. *Automated segmentation of Multiple Sclerosis Lesions by model outlier detection*, in "IEEE trans. Med. Ima.", 2001, vol. 20, n$^{o}$ 8, p. 677-688.