IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble 1)

Activity Report 2011

Project-Team MISTIS

Modelling and Inference of Complex and Structured Stochastic Systems

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Optimization, Learning and Statistical Methods

Table of contents

Project-Team MISTIS

Keywords: Statistical Methods, Mixture Models, Markovian Model

1. Members

Research Scientists
Florence Forbes [Team Leader, DR,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]

Technical Staff
Senan James Doyle [INRIA]
Ludovic Leau-Mercier [INRIA]

PhD Students
Lamiae Azizi [INRA, co-advised by F. Forbes and M. Charras-Garrido (INRA Theix)]
Jonathan El-Methni [INRIA, from October 2010, co-advised by L. Gardes and S. Girard]
El-Hadji Deme [Université Gaston Berger, Sénégal]
Christine Bakhous [INRIA, from November 2010, co-advised by F. Forbes and M. Dojat (GIN)]
Gildas Mazo [INRIA, from October 2011, co-advised by F. Forbes and S. Girard]

Post-Doctoral Fellows
Darren Wraith [INRIA]
Kai Qin [INRIA]
Laure Amate [CNRS, until August 2011]
Lotfi Chaari [INRIA]
Huu Giao Nguyen [INRIA, since November 2011]
Farida Enikeeva [INRIA]
Thomas Vincent [INRIA, since Dec. 2011]

Administrative Assistant
Imma Presseguer [INRIA]

Others
Federico Raimondo [INRIA, July-Dec. 2011]
Hessam Hessami [INRIA, May-July 2011]

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

2.2.1. Outstanding paper award at ICMI'11

Our article "Finding Audio-Visual Events in Informal Social Gatherings" [21] received the "Outstanding Paper Award" (best paper) at the IEEE/ACM 13th International Conference on Multimodal Interaction (ICMI), Alicante, Spain, November 2011. The paper is co-authored by members of both PERCEPTION and MISTIS, Xavi Alameda-Pineda, Vasil Khalidov, Radu Horaud and Florence Forbes. The paper addresses the problem of detecting and localizing audio-visual events (such as people) in a complex/cluttered scenario such as a cocktail party. The work is carried out within the collaborative European project HUMAVIPS.
BEST PAPER AWARD :
[21] **IEEE/ACM International Conference on Multimodal Interfaces**. X. ALAMEDA-PINEDA, V. KHALIDOV, R. HORAUD, F. FORBES.

3. Scientific Foundations

3.1. Mixture models

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

In a first approach, we consider statistical parametric models, θ being the parameter, possibly multi-dimensional, usually unknown and to be estimated. We consider cases where the data naturally divides into observed data $y = y_1, ..., y_n$ and unobserved or missing data $z = z_1, ..., z_n$. The missing data z_i represents for instance the memberships of one of a set of K alternative categories. The distribution of an observed y_i can be written as a finite mixture of distributions,

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

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

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

3.2. Markov models

Participants: Laure Amate, Lamiae Azizi, Christine Bakhous, Lotfi Chaari, Senan James Doyle, Jean-Baptiste Durand, Florence Forbes, 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: El-Hadji Deme, Jonathan El-Methni, Laurent Gardes, Stéphane Girard, Gildas Mazo, Kai Qin, Huu Giao Nguyen, Farida Enikeeva.

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 [58]. 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 [63] which combines non-parametric regression techniques with

parametric dimension reduction aspects. This is also the case in *extreme value analysis* [57], which is based on the modelling of distribution tails (see section 3.3.1). It differs from traditional statistics which focuses on the central part of distributions, *i.e.* on the most probable events. Extreme value theory shows that distribution tails can be modelled by both a functional part and a real parameter, the extreme value index.

3.3.1. Modelling extremal events

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

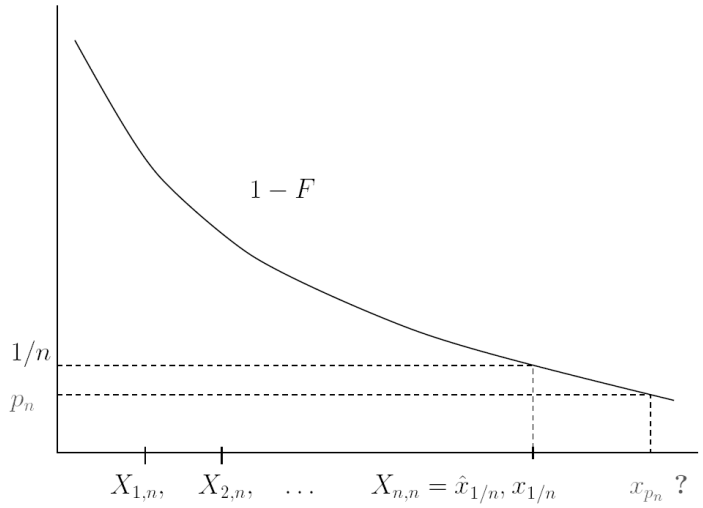

Figure 1. The curve represents the survival function $x \to P(X > x)$. The $1/n$-quantile is estimated by the maximum observation so that $\widehat{x}_{1/n} = X_{n,n}$. As illustrated in the figure, to estimate p_n-quantiles with $p_n < 1/n$, it is necessary to extrapolate beyond the maximum observation.

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

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

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

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

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

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

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

3.3.2. Level sets estimation

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

3.3.3. Dimension reduction

Our work on high dimensional data requires that we face the curse of dimensionality phenomenon. Indeed, the modelling of high dimensional data requires complex models and thus the estimation of high number of parameters compared to the sample size. In this framework, dimension reduction methods aim at replacing the original variables by a small number of linear combinations with as small as a possible loss of information. Principal Component Analysis (PCA) is the most widely used method to reduce dimension in data. However, standard linear PCA can be quite inefficient on image data where even simple image distorsions can lead to highly non-linear data. Two directions are investigated. First, non-linear PCAs can be proposed, leading to semi-parametric dimension reduction methods [60]. Another field of investigation is to take into account the application goal in the dimension reduction step. One of our approaches is therefore to develop new Gaussian models of high dimensional data for parametric inference [53]. Such models can then be used in a Mixtures or Markov framework for classification purposes. Another approach consists in combining dimension reduction, regularization techniques, and regression techniques to improve the Sliced Inverse Regression method [63].

4. Software

4.1. The ECMPR software

Participant: Florence Forbes.

Joint work with: Radu Horaud and Manuel Iguel.

The ECMPR (Expectation Conditional Maximization for Point Registration) package implements [56] [17]. 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.

4.2. The LOCUS and P-LOCUS software

Participants: Florence Forbes, Senan James Doyle.

Joint work with: Michel Dojat.

From brain MR images, neuroradiologists are able to delineate tissues such as grey matter and structures such as Thalamus and damaged regions. This delineation is a common task for an expert but unsupervised segmentation is difficult due to a number of artefacts. The LOCUS software and its recent extension P-LOCUS automatically perform this segmentation for healthy and pathological brains An image is divided into cubes on each of which a statistical model is applied. This provides a number of local treatments that are then integrated to ensure consistency at a global level, resulting in low sensitivity to artifacts. The statistical model is based on a Markovian approach that enables to capture the relations between tissues and structures, to integrate a priori anatomical knowledge and to handle local estimations and spatial correlations.

The LOCUS software has been developed in the context of a collaboration between Mistis, a computer science team (Magma, LIG) and a Neuroscience methodological team (the Neuroimaging team from Grenoble Institut of Neurosciences, INSERM). This collaboration resulted over the period 2006-2008 into the PhD thesis of B. Scherrer (advised by C. Garbay and M. Dojat) and in a number of publications. In particular, B. Scherrer received a "Young Investigator Award" at the 2008 MICCAI conference. Its extension for lesion detection is realized by S. Doyle with financial support from Gravit for possible industrial transfer.

The originality of this work comes from the successful combination of the teams respective strengths i.e. expertise in distributed computing, in neuroimaging data processing and in statistical methods.

4.3. The POPEYE software

Participant: Florence Forbes.

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

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

4.4. 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 [53]. The HDDA and HDDC toolboxes are available for Matlab and are included into the software MixMod [52]. 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/.

4.5. 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 [54]. 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/.

4.6. 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. The INRA Toulouse unit is more recently participating to this project for promotion among the bioinformatics community [20].

4.7. 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 [65]. 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 [55].

4.8. 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 [59]. 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

5. New Results

5.1. Mixture models

5.1.1. Taking into account the curse of dimensionality

Participant: Stéphane Girard.

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

In the PhD work of Charles Bouveyron (co-advised by Cordelia Schmid from the INRIA LEAR team) [53], we propose new Gaussian models of high dimensional data for classification purposes. We assume that the data live in several groups located in subspaces of lower dimensions. Two different strategies arise:

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

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

In collaboration with Gilles Celeux and Charles Bouveyron, we have designed an automatic selection of the discrete parameters of the model [12]. Also, the description of the R package is submitted for publication [44].

5.1.2. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: Application to robust clustering

Participants: Florence Forbes, Darren Wraith.

We proposed a family of multivariate heavy-tailed distributions that allow variable marginal amounts of tailweight. The originality comes from the eigenvalue decomposition of the covariance matrix in the traditional Gaussian scale mixture representation. By contrast to most existing approaches, the derived distributions can account for a variety of shapes and have a simple tractable form with a closed-form probability density function whatever the dimension. We examined a number of properties of these distributions and illustrate them in the particular case of Pearson type VII and t tails. For these latter cases, we provided maximum likelihood estimation of the parameters and illustrated their modelling flexibility on clustering examples for several simulated and real data sets.

5.2. Markov models

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

Participants: Florence Forbes, Lotfi Chaari, Thomas Vincent.

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

In standard fMRI within-subject analysis, two steps are generally performed separately: detection and estimation. Because these two steps are inherently linked, we proposed in this work a joint detection-estimation procedure. We adopt the so-called region-based Joint Detection Estimation (JDE) framework that deals with spatial dependencies between voxels belonging to the same functionally homogeneous *parcel* in the mask of the 3D brain. After building a spatially adaptive General Linear Model, prior information is introduced and a hierarchical Bayesian model is established. In contrast to previous works that use Markov Chain Monte Carlo (MCMC) techniques to approximate the resulting intractable posterior distribution, we recast the JDE into a missing data framework and derive a Variational Expectation-Maximization (VEM) algorithm for its inference. It follows a new algorithm that exhibits interesting properties compared to the previously used MCMC-based approach. Experiments on artificial and real data show that VEM-JDE is robust to model mis-specification and provides computational gain while maintaining good performance. Corresponding papers [27], [38], [26].

5.2.2. Adaptive experimental condition selection in event-related fMRI

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

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

Standard Bayesian analysis of event-related functional Magnetic Resonance Imaging (fMRI) data usually assumes that all delivered stimuli possibly generate a BOLD response everywhere in the brain although activation is likely to be induced by only some of them in specific brain areas. Criteria are not always available to select the relevant conditions or stimulus types (e.g. visual, auditory, etc.) prior to estimation and the unnecessary inclusion of the corresponding events may degrade the results. To face this issue, we propose within a Joint Detection Estimation (JDE) framework, a procedure that automatically selects the conditions according to the brain activity they elicit. It follows an improved activation detection that we illustrate on real data.

5.2.3. Finding Audio-Visual Events in Informal Social Gatherings

Participant: Florence Forbes.

Joint work with: Xavier Alameida-Pineda and Radu Horaud from the INRIA Perception team.

In this work [21] we addressed the problem of detecting and localizing objects that can be both seen and heard, e.g., people. This may be solved within the framework of data clustering. We proposed a new multimodal clustering algorithm based on a Gaussian mixture model, where one of the modalities (visual data) is used to supervise the clustering process. This was made possible by mapping both modalities into the same metric space. To this end, we fully exploited the geometric and physical properties of an audio-visual sensor based on binocular vision and binaural hearing. We proposed an EM algorithm that is theoretically well justified, intuitive, and extremely efficient from a computational point of view. This efficiency makes the method implementable on advanced platforms such as humanoid robots. We described in detail tests and experiments performed with publicly available data sets that yield very interesting results.

5.2.4. 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 and Myriam Garrido from INRA Clermont-Ferrand-Theix.

We recast the disease mapping issue of automatically classifying geographical units into risk classes as a clustering task using a discrete hidden Markov model and Poisson class-dependent distributions. The designed hidden Markov prior is non standard and consists of a variation of the Potts model where the interaction parameter can depend on the risk classes. The model parameters are estimated using an EM algorithm and the mean field approximation. This provides a way to face the intractability of the standard EM in this spatial context, with a computationally efficient alternative to more intensive simulation based Monte Carlo Markov Chain (MCMC) procedures. We then focus on the issue of dealing with very low risk values and small numbers of observed cases and population sizes. We address the problem of finding good initial parameter values in this context and develop a new initialization strategy appropriate for spatial Poisson mixtures in the case of not so well separated classes as encountered in animal disease risk analysis. Using both simulated and real data, we compare this strategy to other standard strategies and show that it performs well in a lot of situations. Corresponding papers and communications [43], [24], [37], [25].

5.2.5. Probabilistic model definition for physiological state monitoring

Participants: Laure Amate, Florence Forbes.

Joint work with: Catherine Garbay, Julie Fontecave-Jallon and Benoit Vettier from LIG.

Assessing the global situation of a person from physiological data is a well-known difficult problem. In previous work, we proposed a system that does not produce a diagnosis but instead follows a set of hypotheses and decides of an alarming situation with this information. In this work [22], we focus on data processing part of the system taking into account the complexity and the ambiguity of the data. We propose a statistical approach with a global model based on Hidden Markov Model and we present data models that rely on classical physiological parameters and expert's knowledge. We then learn a model that depends on the person and its environment, and we define and compute confidence values to assess the plausibility of hypotheses.

5.2.6. Solder Paste Inspection

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

This is joint work with VI-Technology.

The majority of defects in PCB manufacture are attributed to the stencil printing process. Stencil printing is the process where *solder paste bricks* are deposited on the PCB *pads*. Solder paste deposition is required to be accurate and repeatable, however complex physical process make this problematic. Components are placed, and their leads are pushed into the solder paste. The solder paste is then melted using, for example, *reflow soldering*.

Inspection can be performed before the solder paste is melted, and it is more economical to identify defects at this stage.
The evaluation of solder paste joint quality involves the analysis of a number of indicative measurements. From these measurements, potential faults are identified and inspected manually. The general challenge is to reduce of the number of potential faults by better analyzing the indicative factor measurements. That is, to improve the *first pass yield* (FPY) which is the percentage of total solder deposits that are good, and that do not require manual inspection. However, the ability to catch defects must be retained. Another aspect to consider is the temporal nature of the process; The mechanism for identifying faults needs to be retrained after a period of time, and so a solution must be capable of using a small training dataset.
It is important to understand and identify the factors that influence quality. The industry standard factor for measuring quality is solder volume. The precise volume is not directly observable, and so is estimated. Often, height is used as a proxy measure for solder bricks of equal area and shape. There are many other contributing factors, however not all of these can be measured directly, making accurate quality determination difficult.

Stencil printing process control attempts to adjust machine parameters according to informative factors. Online printing process control faces a similar challenge of using a limited number of measurements to inform on the quality of solder paste deposition.
We used statistical techniques to analyze such measurements. The exact nature of the work is confidential.

5.2.7. PCB defect detection

Participants: Florence Forbes, Kai Qin, Huu Giao Nguyen.

This is joint work with VI-Technology.

The objective is to detect defective components in PC Boards from image data. The exact nature of the work is confidential.

5.2.8. Statistical characterization of tree structures based on Markov Tree Models and multitype branching processes, with applications to tree growth modeling.

Participant: Jean-Baptiste Durand.

Joint work with: Pierre Fernique (Montpellier 2 University and CIRAD) and Yann Guédon (CIRAD), INRIA Virtual Plants.

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

Since the connections between shoots can be represented by mathematical trees, statistical models based on multitype branching processes and Markov trees appear as a natural tool to model the dependencies of interest. Formally, the properties of a vertex are summed up using the notion of vertex state. In such models, the numbers of children in each state given the parent state are modelled through discrete multivariate distributions. Model selection procedures are necessary to specify parsimonious distributions. We developed an approach based on probabilistic graphical models to identify and exploit properties of conditional independence between numbers of children in different states, so as to simplify the specification of their joint distribution. The graph building stage was based on a Poissonian Generalized Linear Model for the contingency tables of the counts of joint children state configurations. Then, parametric families of distributions were implemented and compared statistically to provide probabilistic models compatible with the estimated independence graph.

This work was carried out in the context of Pierre Fernique's Master 2 internship (Montpellier 2 University and AgroParisTech). It was applied to model dependencies between short or long, vegetative or flowering shoots in apple trees. The results highlighted contrasted patterns related to the parent shoot state, with interpretation in terms of alternation of flowering (see paragraph 5.2.9. This work will be continued during Pierre Fernique's PhD thesis, with extensions to other fruit tree species and other strategies to build probabilistic graphical models and parametric discrete multivariate distributions including covariates and mixed effects.

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

Participant: Jean-Baptiste Durand.

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

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

To model alternation of flowering at AS scale, a second-order Markov tree model was built. The ASs were of two types: flowering or vegetative. Generalized Linear Mixed Models (GLMMs) were used to model the effet of year, replications and genotypes (with their interactions with year or memories of the Markov model) on the transition probabilities. This work was the continuation of the Master 2 internship of Jean Peyhardi (Bordeaux 2 University) and was carried out in the context of the PhD thesis of Baptiste Guitton.

This PhD thesis also comprised the study of alternation in flowering at individual scale, with annual time step. To relate alternation of flowering at AS and individual scales, indices were proposed to characterize alternation at individual scale. The difficulty is related to early detection of alternating genotypes, in a context where alternation is often concealed by a substantial increase of the number of flowers over consecutive years. To separate correctly the increase of the number of flowers due to aging of young trees from alternation in flowering, our model relied on a parametric hypothesis on the base effect random slopes specific to genotype and replications), which translated into mixed effect modelling. Different indices of alternation were then computed on the residuals. Clusters of individuals with contrasted patterns of bearing habits were identified. Our models highlighted significant correlations between indices of alternation at AS and individual scales. The roles of local alternation and asynchronism in regularity of flowering were assessed using an entropy-based criterion, which characterized asynchronism.

As a perspective of this work, patterns in the production of children ASs (numbers of flowering and vegetative children) depending on the type of the parent AS must be analyzed using branching processes and different types of Markov trees, in the context of Pierre Fernique's PhD Thesis (see paragraph 5.2.8).

5.3. Semi and non-parametric methods

5.3.1. Harmony Search with Differential Mutation Based Pitch Adjustment

Participants: Kai Qin, Florence Forbes.

Harmony search (HS), as an emerging metaheuristic technique mimicking the improvisation behavior of musicians, has demonstrated strong efficacy of solving various numerical and real-world optimization problems. This work [36] presents a harmony search with differential mutation based pitch adjustment (HSDM) algorithm, which improves the original pitch adjustment operator of HS using the self-referential differential mutation scheme that features differential evolution - another celebrated metaheuristic algorithm. In HSDM, the differential mutation based pitch adjustment can dynamically adapt the properties of the landscapes being explored at different searching stages. Meanwhile, the pitch adjustment operator's execution probability is allowed to vary randomly between 0 and 1, which can maintain both wild and fine exploitation throughout the searching course. HSDM has been evaluated and compared to the original HS and two recent HS variants using 16 numerical test problems of various searching landscape complexities at 10 and 30 dimensions. HSDM consistently demonstrates superiority on most of test problems.

5.3.2. Dynamic Regional Harmony Search Algorithm with Opposition and Local Learning

Participants: Kai Qin, Florence Forbes.

To deal with the deficiencies associated with the original Harmony Search (HS) such as premature convergence and stagnation, a dynamic regional harmony search (DRHS) algorithm incorporating opposition and local learning is proposed [35]. DRHS utilizes the opposition-based initialization, and performs independent HS with respect to multiple groups that are randomly recreated on a fixed period basis. Besides the traditional harmony improvisation operators, an opposition based harmony creation scheme is introduced to update the group memory. Any prematurely converged group will be restarted with the doubled size to further augment its exploration capability. Local search is periodically applied to exploit promising regions around top-ranked candidate solutions. The performance of DRHS has been evaluated and compared to HS using 12 numerical test problems at 10D and 30D, which are taken from the CEC2005 benchmark. DRHS consistently demonstrate superiority to HR over all the test problems at both 10D and 30D.

5.3.3. Evolutionary algorithms with CUDA

Participants: Kai Qin, Federico Raimondo.

Evolutionary algorithms (EAs), inspired by natural evolution processes, have demonstrated strong efficacy for solving various real-world optimization problems, although their practical use may be constrained by their computation efficiency. In fact, EAs are inherently parallelizable due to the operations at the individual element level and population-wise evolution. However, most of the existing EAs are designed and implemented in the sequential manner mainly because hardware platforms supporting parallel computing tasks and software platforms facilitating parallel programming tasks are not prevalently available.

In recent year, the graphics processing unit (GPU) has emerged as a powerful general-purpose computation device that can favorably support massively data parallel computing tasks carried out on its hundreds of cores. The compute unified device architecture (CUDA) technology invented by NVIDIA provides an intuitive way to express parallelism and to implement parallel programs using some popular programming languages, such as C, C++ and FORTRAN. Accordingly, we can simply write a program for one data elements, which gets automatically distributed across hundreds of cores for thousands of threads to execute. Although the CUDA programming model is easy-to-use, the computation efficiency of CUDA parallel programs crucially depends on careful consideration of hardware characteristics of GPUs during algorithmic design and implementation, especially about memory utilization and thread management (to maximize the occupancy of streaming multi-processors). Without proper considerations, the parallel programs may even run slower than their sequential counterparts.

The objectives of our project are to: 1. Redesign state-of-the-art EAs using CUDA under thorough consideration of GPU's hardware characteristics. 2. Develop a generic hardware-self-configurable EA framework, which allows automatically configuring available hardware computing resources to maximize the computation efficiency of the EA.

Currently, we had developed a memory-efficient parallel differential evolution algorithm, which features maximally utilizing the available shared memory in GPU while maximally reducing the use of the global memory in GPU considering its very limited access bandwidth. Compared with two recent parallel differential evolution algorithms implemented with CUDA in 2010 and 2011, our algorithm demonstrated significantly faster computation speed. We had also investigated the parallel implementation of test problems and provided a guideline on how to implement any user-defined test problem and combine it with an existing parallel EA framework. To the best of our knowledge, this is the first research work on this topic.

5.3.4. 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 $\theta > 0$ [16]. This model includes very different distribution tail behaviors from Fréchet and Gumbel maximum domains of attraction. In the particular cases of Pareto type tails ($\tau = 1$) or Weibull tails ($\tau = 0$), our estimators coincide with classical ones proposed in the literature, thus permitting us to retrieve their asymptotic normality in an unified way. The first year of the PhD work of Jonathan El-methni has been dedicated to the definition of an estimator of the parameter τ. This permits the construction of new estimators of extreme quantiles. The results are submitted for publication [48]. Our future work will consist in proposing a test procedure in order to discriminate between Pareto and Weibull tails.
We are also working on the estimation of the second order parameter ρ (see paragraph 3.3.1). We proposed a new family of estimators encompassing the existing ones (see for instance [62], [61]). This work is in collaboration with El-Hadji Deme, a PhD student from the Université de Saint-Louis (Sénégal). El-Hadji Deme obtained a one-year mobility grant to work within the Mistis team on extreme-value statistics. The results are submitted for publication [46].

5.3.5. *Conditional extremal events*

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

Joint work with: J. Carreau, A. Lekina, 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 [58] with extreme-value methods in order to obtain efficient estimators of the conditional tail index and conditional extreme quantiles. When the covariate is random (random design) and the tail of the distribution is heavy, we focus on kernel methods [14]. We extension to all kind of tails in investigated in [45].

Conditional extremes are studied in climatology where one is interested in how climate change over years might affect extreme temperatures or rainfalls. In this case, the covariate is univariate (time). Bivariate examples include the study of extreme rainfalls as a function of the geographical location. The application part of the study is joint work with the LTHE (Laboratoire d'étude des Transferts en Hydrologie et Environnement) located in Grenoble.

More future work will include the study of multivariate and spatial extreme values. With this aim, a research on some particular copulas [1] has been initiated with Cécile Amblard, since they are the key tool for building multivariate distributions [64]. The PhD theses of Jonathan El-methni and Gildas Mazo should address this issue too.

5.3.6. *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 [41]: 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 [51].

5.3.7. Quantifying uncertainties on extreme rainfall estimations

Participants: Laurent Gardes, Stéphane Girard.

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

Extreme rainfalls are generally associated with two different precipitation regimes. Extreme cumulated rainfall over 24 hours results from stratiform clouds on which the relief forcing is of primary importance. Extreme rainfall rates are defined as rainfall rates with low probability of occurrence, typically with higher mean return-levels than the maximum observed level. For example Figure 2 presents the return levels for the Cévennes-Vivarais region obtained in [14]. It is then of primary importance to study the sensitivity of the extreme rainfall estimation to the estimation method considered.

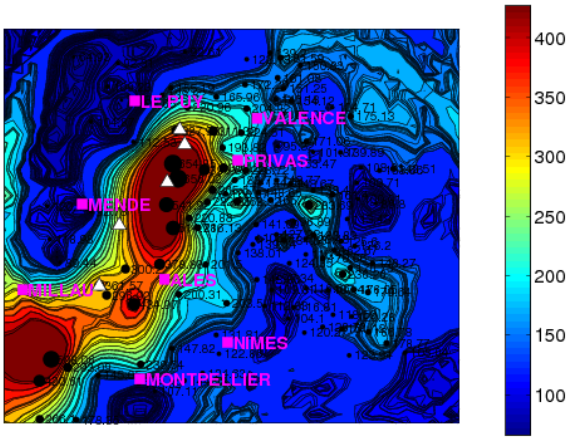

Figure 2. Map of the mean return-levels (in mm) for a period of 10 years.

The obtained results are published in [13].

5.3.8. Retrieval of Mars surface physical properties from OMEGA hyperspectral images.

Participant: Stéphane Girard.

Joint work with: Douté, S. from Laboratoire de Planétologie de Grenoble, France and Saracco, J (University Bordeaux).

Visible and near infrared imaging spectroscopy is one of the key techniques to detect, to map and to characterize mineral and volatile (eg. water-ice) species existing at the surface of planets. Indeed the chemical composition, granularity, texture, physical state, etc. of the materials determine the existence and morphology of the absorption bands. The resulting spectra contain therefore very useful information. Current imaging spectrometers provide data organized as three dimensional hyperspectral images: two spatial dimensions and one spectral dimension. Our goal is to estimate the functional relationship F between some observed spectra and some physical parameters. To this end, a database of synthetic spectra is generated by a physical radiative transfer model and used to estimate F. The high dimension of spectra is reduced by Gaussian regularized

sliced inverse regression (GRSIR) to overcome the curse of dimensionality and consequently the sensitivity of the inversion to noise (ill-conditioned problems). We have also defined an adaptive version of the method which is able to deal with block-wise evolving data streams [28].

5.3.9. Statistical modelling development for low power processor.

Participant: Stéphane Girard.

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

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

The need for an effective modeling of process variation for static power analysis has led to the introduction of Statistical static power analysis. Some publication state that it is possible to save up to 50% static power using statistical approach. However, most of the statistical approaches are based on Monte Carlo analysis, and such methods are not suited to large devices. It is thus necessary to develop solutions for large devices integrated in an industrial design flow. Our objective to model the total consumption of the circuit from the probability distribution of consumption of each individual gate. Our preliminary results are published in [18].

6. Partnerships and Cooperations

6.1. National Actions

MISTIS is a partner in a 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. The goal is to address variability issues when designing electronic components.

MISTIS got, for the period 2008-2011, 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/).

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

6.2. Regional Initiatives

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

6.3. European Initiatives

6.3.1. FP7 Projet

6.3.1.1. HUMAVIPS

Title: Humanoids with audiovisual skills in populated spaces

Type: COOPERATION (ICT)

Defi: Cognitive Systems and Robotics

Instrument: Specific Targeted Research Project (STREP)

Duration: February 2010 - January 2013

Coordinator: INRIA (France)

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

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

Abstract: Humanoids expected to collaborate with people should be able to interact with them in the most natural way. This involves significant perceptual, communication, and motor processes, operating in a coordinated fashion. Consider a social gathering scenario where a humanoid is expected to possess certain social skills. It should be able to explore a populated space, to localize people and to determine their status, to decide to join one or two persons, to synthetize appropriate behavior, and to engage in dialog with them. Humans appear to solve these tasks routinely by integrating the often complementary information provided by multi sensory data processing, from low-level 3D object positioning to high-level gesture recognition and dialog handling. Understanding the world from unrestricted sensorial data, recognizing people's intentions and behaving like them are extremely challenging problems. The objective of HUMAVIPS is to endow humanoid robots with audiovisual (AV) abilities: exploration, recognition, and interaction, such that they exhibit adequate behavior when dealing with a group of people. Proposed research and technological developments will emphasize the role played by multimodal perception within principled models of human-robot interaction and of humanoid behavior. An adequate architecture will implement auditory and visual skills onto a fully programmable humanoid robot. An open-source software platform will be developed to foster dissemination and to ensure exploitation beyond the lifetime of the project. The MISTIS contribution will consist in developing statistical machine learning techniques for interactive robotic applications.

6.4. International Initiatives

6.4.1. Visits of International Scientists

6.4.1.1. Internships

Federico Raimondo (from Jul 2011 until Dec 2011)

Subject: Parallel Self-Adaptive Evolutionary Optimization Framework on GPU
Institution: Universidad de Buenos Aires (Argentina)

El Hadji DEME (from Apr 2011 until Dec 2011)

Subject: Estimation de copules extremaux, de la densite spectrale multivariee et applications : Biologie et changements climatiques
Institution: Universite Gaston Berger (Senegal)

7. Dissemination

7.1. Animation of the scientific community

Florence Forbes and Stéphane Girard co-organized the workshops “Astrostatistique en France” http://astrostat.sciencesconf.org/ and Statlearn, "Challenging problems in Statistical Learning" http://mistis.inrialpes.fr/statlearn/ in Grenoble.

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. In this context, she organized with R. Munos, B. Espiau and M. Thonnat an INRIA workshop on Statistical Learning in Paris (December).

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 Paris I.

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) to evaluate research proposals,
- evaluation of interdisciplinary and inter-institutes projects (PEPII) for the CNRS,
- the national fund for research of Québec - Nature and technology (FRQNT) to evaluate research proposals.

S. Girard was involved in the following PhD committees

- Mohammed El Anbari *“Regularisation and variable selection using penalized likelihood”*, Paris-Sud University and Cadi Ayyad University, december 2011.
- Dmitri Novikov “*Statistical methods of detection of current flow structures in stretches of water*”, Montpellier University, december 2011.
- Davide Ceresetti “*Structure spatio-temporelle des fortes précipitations: application à la région Cévennes-Vivarais.*”, Grenoble University, january 2011.

F. Forbes was involved in the PhD committes of Flora Jay from TimB, Univ. Grenoble I. PhD title:"Méthodes bayésiennes pour la génétique des populations: relations entre structure génétique des populations et environnement" (October 2011).

F. Forbes was also involved in the HDR committee of Cécile Hardouin, assistant professor at Paris Ouest Nanterre La Défense University (July 2011). Title:"Quelques contributions à la modélisation et l'analyse statistique de processus spatiaux".

F. Forbes was also involved in the Master committee of Arun Shivanandan from IBIS team(June 2011). Title: Stochastic modelling and indentification of arabinose uptake network in Escherichia coli.

7.2. Teaching

Stéphane Girard

Master : Statistique inférentielle avancée, 27h, M1, Ensimag (Grenoble INP), France.

Master : Statistique des valeurs extrêmes, 45h, M2, Université Grenoble I, France.

Florence Forbes

Master : Mixture models and EM algorithm, 12h, M2, UFR IM2A, Université Grenoble I, France.

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

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

PhD & HdR :

PhD : Lamiae Azizi, Champs aléatoires de Markov cachés pour la cartographie du risque en épidémiologie, Université Joseph Fourier, December 13, Florence Forbes and Myriam Garrido

PhD in progress : Jonathan El Methni, Différentes contributions à l'estimation des quantiles extrêmes, October, 2010, Stéphane Girard et Laurent Gardes

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

PhD in progress : Gildas Mazo, Estimation de quantiles extrêmes spatiaux, October, 2011, Florence Forbes and Stéphane Girard

8. 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° 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° 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] L. AZIZI. *Champs aléatoires de Markov cachés pour la cartographie du risque en épidémiologie*, Université Joseph-Fourier - Grenoble I, December 2011.

Articles in International Peer-Reviewed Journal

[12] C. BOUVEYRON, G. CELEUX, S. GIRARD. *Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA*, in "Pattern Recognition Letters", 2011, vol. 32, p. 1706-1713 [*DOI :* 10.1016/J.PATREC.2011.07.017], http://hal.inria.fr/hal-00440372/en.

[13] J. CARREAU, S. GIRARD. *Spatial extreme quantile estimation using a weighted log-likelihood approach*, in "Journal de la Société Française de Statistique", 2011, vol. 152, n° 3, p. 66–83.

[14] A. DAOUIA, L. GARDES, S. GIRARD, A. LEKINA. *Kernel estimators of extreme level curves*, in "Test", 2011, vol. 20, n° 14, p. 311–333.

[15] F. FORBES, B. SCHERRER, M. DOJAT. *Bayesian Markov model for cooperative clustering: application to robust MRI brain scan segmentation*, in "Journal de la Societe Française de Statistique", 2011, vol. 152, n° 3.

[16] L. GARDES, S. GIRARD, A. GUILLOU. *Weibull tail-distributions revisited: a new look at some tail estimators*, in "Journal of Statistical Planning and Inference", 2011, vol. 141, n° 1, p. 429–444.

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

[18] S. JOSHI, A. LOMBARDOT, P. FLATRESSE, C. D'AGOSTINO, A. JUGE, E. BEIGNE, S. GIRARD. *Statistical estimation of dominant physical parameters for leakage variability in 32nanometer CMOS under supply voltage variations*, in "Journal of Low Power Electronics", 2011, vol. 7, n^{o} 5, to appear.

[19] V. KHALIDOV, F. FORBES, R. HORAUD. *Conjugate Mixture Models for Clustering Multimodal Data*, in "Neural Computation", February 2011, vol. 23, n^{o} 2, p. 517–557 [*DOI :* 10.1162/NECO_a_00074], http://hal.inria.fr/inria-00590267/en.

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

International Conferences with Proceedings

[21] *Best Paper*
X. ALAMEDA-PINEDA, V. KHALIDOV, R. HORAUD, F. FORBES. *Finding Audio-Visual Events in Informal Social Gatherings*, in "IEEE/ACM International Conference on Multimodal Interfaces", Alicante, Spain, November 2011, http://hal.inria.fr/inria-00623489/en.

[22] L. AMATE, F. FORBES, J. FONTECAVE-JALLON, B. VETTIER, C. GARBAY. *Probabilistic Model Definition for Physiological State Monitoring*, in "IEEE International Workshop on Statistical Signal Processing 2011 (SSP11)", June 28-30 2011.

[23] A. ASENOV, Y. COURANT, G. DUCHARME, V. GEROUSIS, S. GIRARD. *How can statistics methods help to address variability issues?*, in "2nd European Workshop on CMOS Variability", Grenoble, 2011.

[24] L. AZIZI, D. ABRIAL, M. CHARRAS-GARRIDO, F. FORBES. *Risk mapping based on hidden Markov random field and variational approximations*, in "1st Conference on Spatial Statistics 2011 - Mapping Global Change", University of Twente, Enschede, The Netherlands, 2011.

[25] L. AZIZI, F. FORBES, S. DOYLE, M. CHARRAS-GARRIDO, D. ABRIAL. *Spatio-temporal Markov Random Field approach to risk mapping*, in "The 14th Applied Stochastic Models and Data Analysis (ASMDA2011) conference of the ASMDA International Society", Rome, Italy, June 2011.

[26] L. CHAARI, F. FORBES, P. CIUCIU, T. VINCENT, M. DOJAT. *Bayesian Variational Approximation for the Joint Detection Estimation of Brain Activity in fMRI*, in "IEEE International Workshop on Statistical Signal Processing 2011 (SSP11)", June 28-30 2011.

[27] L. CHAARI, F. FORBES, T. VINCENT, M. DOJAT, P. CIUCIU. *Med Image Comput Comput Assist Interv*, 2011, vol. 14, n^{o} Pt 2, p. 260-8, http://hal.inria.fr/inserm-00635384/en.

[28] M. CHAVENT, S. GIRARD, V. KUENTZ, B. LIQUET, T. NGUYEN, J. SARACCO. *An adaptive SIR method for block-wise evolving data streams*, in "XIVth International Symposium of Applied Stochastic Models and Data Analysis (ASMDA 2011)", Rome, Italy, 2011, 8, http://hal.inria.fr/hal-00601924/en.

[29] A. CLÉMENT, S. LAURENS, S. GIRARD. *A Novel Damage Sensitive Feature Based on State-Space Representation*, in "8th International Workshop on Strutural Health Monitoring", Stanford, USA, septembre 2011.

[30] E. DEME, L. GARDES, S. GIRARD. *A new semi-parametric family of estimators for the second order parameter*, in "7th International Conference on Extreme Value Analysis", Lyon, juin 2011.

[31] J. EL-METHNI, L. GARDES, S. GIRARD, A. GUILLOU. *Estimation of a new parameter discriminating between Weibull tail-distributions and heavy-tailed distributions*, in "7th International Conference on Extreme Value Analysis", Lyon, juin 2011.

[32] L. GARDES, S. GIRARD. *Functional kernel estimators of conditional extreme quantiles*, in "2nd International Workshop on Functional and Operatorial Statistics", Santander, Spain, 2011.

[33] S. GIRARD, A. GUILLOU, G. STUPFLER. *Estimating an endpoint using high order moments*, in "7th International Conference on Extreme Value Analysis", Lyon, juin 2011.

[34] S. GIRARD, L. MENNETEAU. *Strong invariance principles for tail quantile processes with applications to extreme value index estimation*, in "7th International Conference on Extreme Value Analysis", Lyon, juin 2011.

[35] K. QIN, F. FORBES. *Dynamic Regional Harmony Search with Opposition and Local Learning*, in "Genetic and Evolutionary Computation Conference 2011 (Gecco 2011)", Dublin, July 2011.

[36] K. QIN, F. FORBES. *Harmony Search with Differential Mutation Based Pitch Adjustment*, in "Genetic and Evolutionary Computation Conference 2011 (Gecco 2011)", Dublin, July 2011.

National Conferences with Proceeding

[37] L. AZIZI, F. FORBES, M. CHARRAS-GARRIDO, D. ABRIAL, S. DOYLE. *Initialisation de l'algorithme EM champ-moyen pour les mélanges de Poisson pour données spatiales et application à la cartographie du risque en épidémiologie*, in "43èmes Journées de Statistique organisées par la Société Française de Statistique", Tunis, Tunisia, May 2011.

[38] L. CHAARI, F. FORBES, P. CIUCIU, T. VINCENT, M. DOJAT. *A Variational Bayesian approach for the Joint Detection Estimation of Brain Activity in functional MRI*, in "43èmes Journées de Statistique organisées par la Société Française de Statistique", Tunis, Tunisia, May 2011.

[39] E. DEME, L. GARDES, S. GIRARD. *Estimation semi-paramétrique du paramètre de second ordre en statistique des valeurs extrêmes*, in "43èmes Journées de Statistique organisées par la Société Française de Statistique", Tunis, mai 2011.

[40] J. EL-METHNI, L. GARDES, S. GIRARD, A. GUILLOU. *Estimation d'un paramètre de queue commun aux lois de type Weibull et au domaine d'attraction de Fréchet*, in "43èmes Journées de Statistique organisées par la Société Française de Statistique", Tunis, mai 2011.

Scientific Books (or Scientific Book chapters)

[41] A. DAOUIA, L. GARDES, S. GIRARD. *Nadaraya's estimates for large quantiles and free disposal support curves*, in "Exploring research frontiers in contemporary statistics and econometrics", I. V. KEILEGOM, P. WILSON (editors), Springer, 2011, to appear.

[42] L. GARDES, S. GIRARD. *Functional kernel estimators of conditional extreme quantiles*, in "Recent advances in functional data analysis and related topics", F. FERRATY (editor), Springer, Physica-Verlag, 2011, p. 135–140.

Research Reports

[43] L. AZIZI, F. FORBES, S. DOYLE, M. CHARRAS-GARRIDO, D. ABRIAL. *Spatial risk mapping for rare disease with hidden Markov fields and variational EM*, INRIA, March 2011, n^{o} RR-7572, http://hal.inria.fr/inria-00577793/en.

Other Publications

[44] L. BERGÉ, C. BOUVEYRON, S. GIRARD. *HDclassif: an R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data*, 2011, http://hal.archives-ouvertes.fr/hal-00541203.

[45] A. DAOUIA, L. GARDES, S. GIRARD. *On kernel smoothing for extremal quantile regression*, 2011, http://hal.inria.fr/hal-00630726/en.

[46] E. DEME, L. GARDES, S. GIRARD. *On the estimation of the second order parameter in extreme-value theory*, 2011, http://hal.inria.fr/hal-00634573/en.

[47] J. DURAND, S. GIRARD, V. CIRIZA, L. DONINI. *Optimization of power consumption and user impact based on point process modeling of the request sequence*, 2011, http://hal.archives-ouvertes.fr/hal-00412509/fr/.

[48] J. EL-METHNI, L. GARDES, S. GIRARD, A. GUILLOU. *Estimation of extreme quantiles from heavy and light tailed distributions*, 2011, http://hal.inria.fr/hal-00627964/en.

[49] L. GARDES, S. GIRARD. *Functional kernel estimators of large conditional quantiles*, 2011, http://hal.inria.fr/hal-00608192/en.

[50] L. GARDES, A. GUILLOU, A. SCHORGEN. *Estimating the conditional tail index by integrating a kernel conditional quantile estimator*, 2011, http://hal.inria.fr/inria-00578479/en.

[51] S. GIRARD, A. GUILLOU, G. STUPFLER. *Frontier estimation with kernel regression on high order moments*, 2011, http://hal.archives-ouvertes.fr/hal-00499369/fr/.

References in notes

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

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

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

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

[56] 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.

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

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

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

[60] 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.

[61] 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.

[62] 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.

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

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

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