INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team MISTIS

Modelling and Inference of Complex and Structured Stochastic Systems

Rhône-Alpes

THEME COG

Activity Report

2005

Table of contents

1. Team

Team leader

Florence Forbes [Research scientist, Inria]

Research scientists

Paulo Gonçalves [Research scientist, Inria]

Project technical staff

Lemine Abdalah [August–October 2005]

Ph. D. students

Juliette Blanchet [MENRT, co-advised by F. Forbes and C. Schmid, Team Lear]
Charles Bouveyron [MENRT, co-advised by S. Girard and C. Schmid, Team Lear]
Hugo Carrão [IGP Lisbon, FCT Portugal, co-advised by P. Gonçalves and M. Caetano, IGP Portugal]
Julien Jacques [Inria until December 2005]
Matthieu Vignes [AC, co-advised by F. Forbes and G. Celeux, Team Select]

Research scientists (partners)

Henri Berthelon [Faculty member, CNAM, Paris]
Gersende Fort [Research scientist, CNRS, Paris]
Laurent Gardes [Faculty member, UPMF, Grenoble]
Stéphane Girard [Faculty member, UJF, Grenoble]

Administrative assistant

Stéphanie Berger

2. Overall Objectives

2.1. Overall Objectives

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

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

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

3. Scientific Foundations

3.1. Mixture models

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

Participants: Juliette Blanchet, Charles Bouveyron, Florence Forbes, Gersende Fort, Paulo Gonçalves, Matthieu Vignes.

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

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

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

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

3.2. Markov models

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

Participants: Juliette Blanchet, Florence Forbes, Gersende Fort, Paulo Gonçalves, Matthieu Vignes.

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

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

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

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

3.3. Functional Inference, semi and non parametric methods

Keywords: *boundary estimation*, *extremes*, *multiresolution analysis*, *non parametric*, *scaling laws*, *singularity spectra*, *wavelets*.

Participants: Laurent Gardes, Stéphane Girard, Paulo Gonçalves.

We also consider methods which do not assume a parametric model. The approaches are non-parametric in the sense that they do not require the assumption of a prior model on the unknown quantities. This property is important since, for image applications for instance, it is very difficult to introduce sufficiently general parametric models because of the wide variety of image contents. As an illustration, the grey-levels surface in an image cannot usually be described through a simple mathematical equation. Projection methods are then a way to decompose the unknown signal or image on a set of functions (*e.g.* wavelets). Kernel methods which rely on smoothing the data using a set of kernels (usually probability distributions), are other examples. Relationships exist between these methods and learning techniques using Support Vector Machine (SVM) as this appears in the context of *boundary estimation* and *image segmentation*. These techniques are also of great use for dimension reduction since they allow to avoid assumptions on the observations distribution. Regarding our use of wavelets, our goal is to perform image fusion between high spatial resolution satellite images and lower resolution image time series sensored at short time periods. Our approach relies on the inherent multiresolution analysis structure of orthogonal wavelets, combined with a hidden Markov tree model to assess the inter-scale statistical dependencies.

4. Application Domains

4.1. Image Analysis

Participants: Juliette Blanchet, Charles Bouveyron, Hugo Carrão, Florence Forbes, Stéphane Girard, Paulo Gonçalves.

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

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

We also consider other statistical 2D fields coming from other domains such as teledetection, remote sensing, or Time-Frequency representations of 1-D signals.

4.2. Biology and Medicine

Participants: Florence Forbes, Matthieu Vignes.

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

4.3. Reliability

Participants: Henri Bertholon, Julien Jacques.

Reliability and industrial lifetime analysis are applications developed essentially through collaborations with the EDF research department and the LCFR laboratory of CEA / Cadarache.

5. Software

5.1. The Extremes freeware

Participant: Stéphane Girard.

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

The EXTREMES software is a toolbox dedicated to the modelling of extremal events offering extreme quantile estimation procedures and model selection methods. This software results from a collaboration with EDF R&D. It is also a consequence of the PhD thesis work of Myriam Garrido. The software is written in C++ with a Matlab graphical interface. It is now available both on Windows and Linux environments. It can be downloaded at the following URL: http://mistis.inrialpes.fr/software/EXTREMES/. Recently, this software has been used to propose a new goodness-of-fit test to the distribution tail [16].

5.2. The Semms package

Participants: Juliette Blanchet, Florence Forbes.

This is joint work with Nathalie Peyrard (INRA Avignon).

The SEMMS (Spatial EM for Markovian Segmentation) program proposes a variety of algorithms for image segmentation using Markov Random Fields. It is mainly based on mean field approximations. The main functionalities of the package include:

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

The package is publicly available at: http://mistis.inrialpes.fr/software/SEMMS.html.

A new package with new functionalities has been written in C++ with the help of Lemine Abdalah, when he was part of our technical staff during summer 2005. This package will be made available in 2006 and will complememt the current SEMMS package.

6. New Results

6.1. Mixture models

6.1.1. Taking into account the curse of dimensionality.

Participants: Charles Bouveyron, Stéphane Girard.

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

In high dimensional spaces, learning methods suffer from the curse of dimensionality: even for large datasets, large parts of the spaces are left empty. In the PhD work of Charles Bouveyron (co-advised by Cordelia Schmid from the INRIA team LEAR, in the ACI Movistar in the “Masse de données” program), we propose new Gaussian models of high dimensional data for classification purposes [13], [48]. We assume that the data live in several groups located in subspaces of lower dimensions. Two different strategies arise:

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

This modelling yields new supervised classification methods called HDDA for High Dimensional Discriminant Analysis. Some versions of this method have been tested on the supervised classification of objects in images. We have developped an adaptation of this approach, named HDDC for High Dimensional Data Clustering, to the unsupervised classification framework. We already, in the context of Juliette Blanchet PhD work (also co-advised with C. Schmid), combined the method to our Markov-model based approach of learning and classification and obtained significant improvement in applications such as texture recognition where the observations are high-dimensional. We also foresee to apply this dimensionality reduction strategy in a remote sensing context, when dealing with multi-temporal and hyper-spectral satellite images (Ph.D. work of Hugo Carrão co-advised by Paulo Gonçalves, see the following section).

We are then also willing to get rid of the Gaussian assumption. To this end, non linear models and semi parametric methods are necessary. Our main project is to adapt the non linear Principal Component Analysis (PCA) method proposed in [22] to the classification problem. This method (first introduced in Stéphane Girard's PhD thesis) relies on the approximation of datasets by manifolds, generalizing the PCA linear subspaces. This approach reveals good performances when data are images [4].

6.1.2. Land cover classification using multi-temporal, hyper-spectral satellite images

Participants: Paulo Gonçalves, Hugo Carrão.

This is joint work with Mário Caetano (IGP, Portugal).

The objective of the present work is to produce a semi-automated land cover classification from multi-spectral and multi-temporal MODIS satellite images acquired at a 500m nominal resolution. Our goal is to achieve an automated pixel level classification using a Support Vector Machine (SVM) learning approach. More specifically, we use the time evolution of reflectances measured in several spectral bands from weekly composited images acquired during a complete year period. As temporal profiles are relevant fingerprints of local phenologies, we believe time series offer great potential to improve discrimination among the different land cover types. However, they result in very high dimensional data that we propose to handle considering two approaches: the first one consists in identifying a parsimonious set of fitting parameters that adequately model the time series. A second approach is based on dimensionality reduction techniques such as principal component analysis and factorial discriminant analysis (see Section above).

Eventually, our model parameters are used as inputs of a supervised SVM classifier. Performance is then exhaustively compared to that obtained when the same classifier is directly applied to a single date multi-spectral reflectance data. First results are reported in [42], [45].

6.2. Markov models

6.2.1. Markov models for the spatial organization of image descriptors.

Participants: Florence Forbes, Juliette Blanchet.

In more and more high-level image analysis, such as feature-based object recognition or object tracking, images are described by local affine-invariant descriptors and by spatial relationships between these descriptors. A graph is associated to an image with the nodes representing feature vectors describing image regions and the edges joining spatially related regions. For tractability, most approaches to recognition assume independence between the features which is an obvious oversimplification. Incorporating information about the spatial organization of the descriptors leads to better recognition results. Current approaches consist in augmenting the data with information coming from the spatial relationships, for instance by using co-occurrence statistics, but without modelling explicitly the dependencies between neighboring descriptors. In such approaches the underlying model is one where the descriptors are statistically independent variables. Our claim is that recognition results can be further improved by considering that descriptors are statistically dependent. We propose to introduce the use of statistical parametric models of the dependence between descriptors. In this work, we chose Hidden Markov Models (HMM) which are both well statistically-based and appropriate models for such a task. They provide parametric models where the parameters have a natural interpretation and can be adjusted to incorporate a priori knowledge with respect to strength of interaction for instance. Their use requires non trivial parameter estimation. We propose to use recent estimation procedures based on the mean field principle of statistical physics and to investigate how to make them accurate and computationally efficient. The particularities of the applications we aim at is the high-dimensionality of the feature vectors (typically 100 dimensional) and the irregularity of the sites at which they are observed. Very few practical optimization techniques are available for such tasks. Such algorithms are usually very sensitive to initialization and may require tuning which may be problematic. By combining an MRF estimation procedure and a dimension reduction technique we show that recognition rates could be improved and that promising results could be obtained using a general statistical formalism. We focused in particular on texture recognition but further work includes other contexts such as object recognition and tracking.

As regards texture recognition (joint work with Cordelia Schmid, LEAR, INRIA Rhône-Alpes), images are described by local affine-invariant descriptors and by spatial relationships between these descriptors. Using sample images, textures are then learned as HMM's and a set of estimated parameters is associated to each texture. At recognition time, another HMM is used to compute, for each feature vector, the membership probabilities to the different texture classes. Preliminary experiments show promising results [29].

6.2.2. Integrated Markov models

Participants: Juliette Blanchet, Florence Forbes, Matthieu Vignes.

By integrated Markov model, we mean specific instance and usage of Markov models that we propose to develop to combine various sources of interactions and information. The models are flexible in that various pairwise relationship information and features of individual data can be easily incorporated. Two features distinguish the integrated approach from other available methods. One is that the integrate approach uses all available sources of information with possibly different weights for different sources of data. The second feature is that as a probabilistic model it provides confidence measures such as posterior probabilities that an object is assigned to a class when used for a classification task.

The novelty we propose is to take into account simultaneously data from individual objects, ie data that make sense and exist for each objects, and data from pairs of objects reflecting for instance some similarity measure defined on the objects. In practice such data can be missing and EM offers a good framework to deal with this case (see [60]).

A wide range of clustering algorithms have been proposed to analyze such data. Approaches fall mainly in two categories. Some focus on individual data and as a consequence assume that they are independent. Others use information on pairs in the form of networks or graphs but do not directly use individual data associated to objects in the networks. Sequential procedures clustering first individual data alone and incorporating additional information only after the clusters are determined are necessarily suboptimal. Our aim is to take into account both type of information in a single procedure. We propose a hidden Markov random field model in which parametric probability distributions will account for the distribution of individual data for each object. Data on pairs will then be included through a graph where the nodes represent the objects and the edges weighted according to pair data, for instance in order to reflect distance or similarity measures between objects. There exist many ways to do that and it is not clear whether they are equivalent in terms of the amount of information taken into account and in terms of clustering results. We applied this approach to genetic data analysis (see below). One of the difficulties is to choose how the various information can be incorporated in the model depending on the goal in mind. This requires a good understanding of the role of each parameter in a Hidden Markov Random Field model. With this in mind, in [58], we investigated the role of *singleton potentials* which are parameters usually ignored in standard Markov model-based segmentation. In [47] we used these potentials to take into account cooperatively two sources of information so that two segmentation processes could refine mutually and lead to better segmentation results (see application to MRI analysis below).

Note that as in the previous section most of this work concerns Markov models on irregular graphs. Choosing the neighborhood structure can then be an additional issue.

Integrated Markov models on irregular grids for clustering gene expression data. Because of the increasingly large amounts of genetic data generated by researchers, there is a great need to develop methodology to analyse and to use the information contained in this data. In this framework, clustering of genes into groups sharing common characteristics becomes a useful exploratory technique.

As an example, one of the most popular tools for exploratory analysis of gene expression data is clustering of genes and/or experiments. Furthermore, clustering is also frequently used as the basis for further computational analysis. For example, the function of a gene can be predicted according to known functions of other genes from the same cluster. More generally, a major challenge in bioinformatics is to reveal interactions between living entities and discover the corresponding biological networks responsible for their biological complexity.

Our aim is to classify biological objects sharing common characteristics so that the resulting clusters could be interpreted. More specifically, we focus on the clustering of genes. A wide range of clustering algorithms have been proposed to analyze gene expression data. As mentionned, we propose a hidden Markov random field model in which parametric probability distributions will account for the distribution of individual data for each gene. Data on pairs, resulting from information in the form of biological networks, will then be included through a graph where the nodes represent the objects and the edges weighted according to pair data, for instance in order to reflect distance or similarity measures between genes. Preliminary investigations are reported in [37].

Distributed and Cooperative Markovian segmentation of both tissues and structures in brain MRI. This is joint work with Benoit Scherrer, Michel Dojat and Christine Garbay from TIMC and INSERM. Accurate tissues and structures segmentation of MRI brain scan is critical for several applications. Markov random fields are commonly used for such a task and require the estimation of the model parameters (Potts model). Some refinements can be introduced into estimation algorithms, but are not sufficient for structure segmentation. We propose [47] to inject anatomical a priori knowledge expressed as fuzzy spatial relations. Knowledge obtained from structure segmentation is also injected in turn into the Markov process of tissues segmentation. Structure and tissue segmentations are thus dynamic and cooperative processes. They are implemented into a multi-agent system, where autonomous entities distributed into the image estimate local Markov fields. We show, using phantoms and real images (acquired on a 3T scanner), that a distributed and cooperative Markov modelling using anatomical knowledge is a powerful approach for MRI brain scan segmentation.

6.2.3. Convergence properties of EM-like algorithms for inference in Hidden Markov Random Fields

Participants: Florence Forbes, Gersende Fort.

For the standard EM algorithm, parameter estimates yield increasing likelihood over the observed data and the convergence behavior of this process is well understood. However, since it is often the case that there are no other feasible choices rather than to resort to the mean field approximation in practical situations, it appears frequently that the mean field approximation is being used to practical problems with little consideration of important issues such as accuracy of the approximation, convergence of the algorithms and so on. As a matter of fact, in the context of Markovian segmentation, theoretical results as regards convergence properties are still missing. Convergence properties of related EM variants (GAM for Generalized Alternating Minimization) have been studied by [57] and [61] but these variants cannot be applied in the MRF segmentation framework and further approximations are required. We are investigating [51] a new algorithm that we proposed, the so-called MCVEM algorithm, which is tractable in practice and for which we prove convergence results. Our algorithm has the advantage on the GAM procedures studied in [57] that it can be applied to perform image segmentation tasks and so on the basis of theoretical convergence results. The basis of our work is the paper [9] which focuses on the convergence properties of the MCEM algorithm. Using similar tools, our key idea is to view the MCVEM algorithm as a stochastic perturbation of a deterministic algorithm, so called VEM, easier to study [57]. Experiments on synthetic and real images show that the algorithm performance is very close and sometimes better to that of [3]. Additional good properties due to its stochastic nature need to be further investigated. This first effective step opens the way to a better understanding of the behavior of a lot of Markov based algorithm.

6.3. Semi and non parametric methods

6.3.1. Boundary estimation

Participants: Stéphane Girard, Laurent Gardes.

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

Boundary estimation, or more generally, level sets estimation is a recurrent problem in statistics which is linked to outlier detection. In biology, one is interested in estimating reference curves, that is to say curves which bound 90% (for example) of the population. Points outside this bound are considered as outliers compared to the reference population. In image analysis, the boundary estimation problem arises in image segmentation as well as in supervised learning. Two different and complementary approaches are developped.

6.3.1.1. Extreme quantiles approach.

Here, the boundary bounding the set of points is viewed as the larger level set of the points distribution. This is then an extreme quantile curve estimation problem. We have proposed estimators based on projection as well as on kernel regression methods applied on the extreme values set [23], [24], for particular set of points. In this specific framework, we have obtained the asymptotic distribution of the estimators. In his PhD work, co-directed by Pierre Jacob and Stéphane Girard, Laurent Gardes [59] has adapted these methods to estimate extreme level sets of non-bounded points distributions.

Our future work will be to define similar methods based on wavelets expansions in order to estimate non-smooth boundaries. Besides, we are also working on the extension of our results to more general sets of points.

6.3.1.2. Linear programming approach.

Here, the boundary of a set of points is defined has a closed curve bounding all the points and with smallest associate surface. It is thus natural to reformulate the boundary estimation method as a linear programming problem [12]. The resulting estimate is parsimonious, it only relies on a small number of points. This method belongs to the Support Vector Machines (SVM) techniques. Their finite sample performances are very impressive but their asymptotic properties are not very well known. We have established the speed of convergence and shown that it is optimal for a particular family of boundaries [21].

6.3.2. Modelling extremal events

Participants: Stéphane Girard, Laurent Gardes.

Joint work with Mhamed El Aroui (ISG, Tunis), Armelle Guillou (Université Paris 6) Myriam Garrido (INRA Clermont-Ferrand), Jean Diebolt (CNRS).

The first part of our work is to propose new estimates of the extremal index. This parameter is important in practice since it drives the behaviour of the distribution tail. The second part is then to deduce estimates for extreme quantiles.

In [19], we investigate the asymptotical behaviour of two new estimates based on double threshold methods.

We also introduce a quasi-conjugate Bayes approach for estimating Generalized Pareto Distribution (GPD) parameters, distribution tails and extreme quantiles within the Peaks-Over-Threshold framework [15]. Bayes credibility intervals are defined, they provide assessment of the quality of the extreme events estimates. Posterior estimates are computed by Gibbs samplers with Hastings-Metropolis steps. Even if non-informative priors are used in this work, the suggested approach could incorporate informative priors. It brings solutions to the problem of estimating extreme events when data are scarce but expert opinion is available.

Finally, we introduce estimates dedicated to the important case of Weibull tail-distributions [20], [52] which includes for instance Gaussian, gamma, and Weibull distributions. Our current work includes kernel estimators [53] and bias reduced estimators [49], [50].

6.3.3. Generalized discriminant rule for binary data when training and test populations differ in their descriptive parameters

Participant: Julien Jacques.

This is a joint work with C. Biernacki, Prof. University of Lilles.

Standard discriminant analysis assumes that both the labelled training sample and the unlabelled test sample which has to be classified both come from the same population. When these samples come from populations for which descriptive parameters are different, generalized discriminant analysis enables us to adapt the classification rule built from the training population to the test population, by estimating a link between these two populations. This work extends existing methods available in a multi-normal context to the case of binary data. To solve the major challenge of this work which is to define a link between the two binary populations, we suppose that binary data come from the discretization of latent Gaussian data. An estimation method is then defined and tests on simulated data are carried out. Also, an application to real biological data illustrates the method [44], [43].

6.3.4. Empirical Mode Decomposition

Participant: Paulo Gonçalves.

This topic is the main line of our scientific collaboration with École Normale Supérieure de Lyon (France). P. Flandrin and P. Gonçalves are co-advising the PhD thesis of G. Rilling (starting date, Sept. 2004) on “Empirical Mode Decomposition" (EMD).

We now briefly describe the EMD technique. This entirely data-driven algorithm introduced by N. E. Huang decomposes iteratively a complex signal (i.e. with several characteristic time scales coexisting) into elementary Amplitude-Frequency Modulation type components (Intrinsic Mode Functions). The rationale of this decomposition is to locally identify in the signal the fastest oscillations, defined as the waveform interpolating interlacing local maxima and minima. To do so, local maxima points (respectively local minima points) are interpolated with a cubic spline, to yield the upper (resp. lower) envelope. The mean envelope (half sum of upper and lower envelopes) is then subtracted from the initial signal, and the same interpolation scheme is re-iterated on the remainder. The so-called *sifting process* stops when the mean envelope is reasonably zero everywhere, and the resulting signal is designated as the first *Intrinsic Mode Function* (IMF). The higher order IMFs are iteratively extracted applying the same procedure to the signal after the previous IMFs have been removed.

We are pursuing the qualitative study of EMD as an adaptive dyadic filter bank. Regarding applications, we are investigating EMD as a tool to estimate the Hurst parameter estimation of fractional Brownian Motions, and show that when compared to wavelet approach [6], EMD yields more accurate estimates, specially for very irregular signals.

6.3.5. Image fusion using Multiresolution Analysis and Markov tree models

Participants: Paulo Gonçalves, Hugo Carrão.

This is joint work with Jean-Baptiste Durand (LMC, Grenoble), and Mário Caetano (IGP, Portugal).

Accurate land cover classification and land cover change estimation from remote sensing require simultaneously fine spatial resolution images and high acquisition time rate. However, sensors able to provide such high quality images are rare and/or very expensive. We propose to cope with this limitation by combining the following two type of images:

1. Images from MODIS sensor. These images have a coarse spatial pixel resolution (250m – 500m) but are periodically acquired at short time intervals (daily or weekly images). They are freely accessible from the NASA Web site.
2. Images from LandSat sensor. These images have high spatial resolution (30m), but long acquisition time period (one year).

The fusion of both sources of information is performed carrying out the following steps. First, the wavelet decomposition of the high resolution LandSat images is computed and the hidden Markov tree model that underlies it is identified according to the work in [5]. This results in a set of Markov transition kernels that can then be applied to the available low resolution Modis images to infer a higher resolution image for each of the date for which no high resolution LandSat image is available. For a given date, the available low resolution Modis image is considered as the wavelet approximation of the non-existent high resolution image at a coarser scale. Applying the learned Markov tree model to it yields a statistical estimate of a higher resolution image for this date.

7. Other Grants and Activities

7.1. Regional initiatives

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

7.2. National initiatives

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

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

7.3. International initiatives

7.3.1. Europe

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

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

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

7.3.2. North Africa

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

7.3.3. North America

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

P. Gonçalves has joint work with:
- M. Caetano (IGP-IGESI Lisbon, Portugal)
- R. Riedi (Rice Univ., USA)
- R. Baraniuk (Rice Univ., USA)
- A. Feuerverger (Univ. of Toronto, CA).

7.4. Visiting scientists

Hugo Carrão (Ph.D. student from IGP, Lisbon Portugal) spent 3 months in the team.

8. Dissemination

8.1. Leadership within scientific community

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

F. Forbes was involved in the PhD commitee of C. Melo de Lima from university Lyon 1.

8.2. University Teaching

F. Forbes lectured a graduate course on the EM algorithm at Univ. J. Fourier, Grenoble and on stochastic processes at ENSIMAG, Telecom, INPG.

L. Gardes, S. Girard are faculty members at Univ. P. Mendes France and Univ. J. Fourier in Grenoble.

H. Berthelon is faculty member at CNAM, Paris.

9. Bibliography

Major publications by the team in recent years

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

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

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

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

[5] J. B. DURAND, P. GONÇALVÈS, Y. GUÉDON. *Statistical Inference for Hidden Markov Tree Models and Application to Wavelet Trees*, in "IEEE, Trans. on Signal Processing", vol. 52, n^{o} 9, 2004, p. 2551–2560.

[6] P. FLANDRIN, P. GONÇALVÈS, P. ABRY. *Lois d'échelle, Fractales et Ondelettes*, Traité Information - Commande - Communication, vol. 2, chap. Analyses en ondelettes et lois d'échelle, Abry, P. and Gonçalvès, P. and Lévy Véhel, J., Paris, France, 2002.

[7] F. FORBES, N. PEYRARD. *Hidden Markov Random Field Model Selection Criteria based on Mean Field-like Approximations*, in "in IEEE trans. PAMI", August 2003.

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

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

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

Articles in refereed journals and book chapters

[11] A. ANTONIADIS, A. FEUERVERGER, GONÇALVÈS. *Wavelet Based Estimation for Univariate Stable Laws*, in "Annals of the Inst. of Stat. Math., Tokyo (JP)", To appear, 2005.

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

[13] C. BOUVEYRON, S. GIRARD, C. SCHMID. *Class-Specific Subspace Discriminant Analysis For High Dimensional Data*, in "Lecture Notes in Computer Science", to appear, 2006.

[14] G. CELEUX, F. FORBES, C. ROBERT, M. TITTERINGTON. *Deviance Information Criteria for missing data models. With discussion*, in "Bayesian Analysis", 2005.

[15] J. DIEBOLT, M. EL-AROUI, M. GARRIDO, S. GIRARD. *Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling*, in "Extremes", to appear, 2005.

[16] J. DIEBOLT, M. GARRIDO, S. GIRARD. *A goodness-of-fit test for the distribution tail*, in "Topics in extreme values, New-York", M. AHSANULLAH, S. KIRMANI (editors). , to appear, Nova Science, 2005.

[17] P. FLANDRIN, P. GONÇALVÈS. *Empirical Mode Decompositions as data-driven wavelet-like expansions*, in "Int. J. of Wavelets, Multiresolution and Information Processing", To appear, 2005.

[18] P. FLANDRIN, P. GONÇALVÈS, G. RILLING. *EMD Equivalent Filter Banks, from Interpretation to Applications*, N. HUANG, S. SHEN (editors). , Interdisciplinary Mathematical Sciences, 2005.

[19] L. GARDES, S. GIRARD. *Asymptotic properties of a Pickands type estimator of the extreme value index*, in "Focus on probability theory, New-York", F. COLOMBUS (editor). , to appear, Nova Science, 2005.

[20] L. GARDES, S. GIRARD. *Estimating extreme quantiles of Weibull tail-distributions*, in "Communication in Statistics - Theory and Methods", vol. 34, n^o 5, 2005, p. 1065–1080.

[21] S. GIRARD, A. IOUDITSKI, A. NAZIN. *L1-optimal frontier estimation via linear programming*, in "Avtomatika i Telemekhanika", vol. 12, 2005, p. 143–161.

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

[23] S. GIRARD, P. JACOB. *Asymptotic normality of the L1-error for Geffroy's estimate of point process boundaries*, in "Publications de l'Institut de Statistique de l'Université de Paris", vol. XLIX, 2005, p. 3–17.

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

[25] P. GONÇALVÈS, R. RIEDI. *Diverging moments and parameter estimation*, in "Journal of American Statistical Association", Also Inria Technical Report RR-4647, 2002, December 2005.

[26] J. GOSME, C. RICHARD, P. GONÇALVÈS. *Adaptive diffusion as a versatile tool for time-frequency and time-scale representations processing: a review*, in "IEEE Trans. on Signal Processing", vol. 53, n^o 11, November 2005, p. 4136–4146.

[27] J. P. OVARLEZ, P. GONÇALVÈS, R. BARANIUK. *Analyse temps-fréquence quadratique III : La classe affine et autres classes covariantes*, F. HLAWATSCH, F. AUGER (editors). , Hermès Sciences Publications, 2005, p. 201–235.

Publications in Conferences and Workshops

[28] J. BLANCHET, F. FORBES, C. SCHMID. *Markov random fields for recognizing textures modeled by feature vectors*, in "International Conference on Applied Stochastic Models and Data Analysis, Brest, France", 2005.

[29] J. BLANCHET, F. FORBES, C. SCHMID. *Markov random fields for textures recognition with local invariant regions and their geometric relationships*, in "British Machine Vision Conference, Oxford, UK", 2005.

[30] J. BLANCHET, F. FORBES, C. SCHMID. *Modèles markoviens pour l'organisation spatiale de descripteurs d'images*, in "7e Conférence francophone sur l'Apprentissage Automatique , Presses Universitaires de Grenoble, Nice, France", 2005, p. 113-126.

[31] J. BLANCHET, F. FORBES, C. SCHMID. *Modèles markoviens pour la reconnaissance de textures à partir de descripteurs locaux et de leur organisation spatiale*, in "37e Journées de Statistique de la Société Francaise de Statistique, Pau, France", 2005.

[32] C. BOUVEYRON, S. GIRARD, C. SCHMID. *Classification of high dimensional data: High dimensional discriminant analysis*, in "Subspace, latent structure and feature selection techniques: statistical and optimisation perspectives workshop, Bohinj, Slovénie", février 2005.

[33] C. BOUVEYRON, S. GIRARD, C. SCHMID. *High dimensional discriminant analysis*, in "International Symposium on Applied Stochastic Models and Data Analysis, ASMDA 2005, Brest", mai 2005, p. 526–534.

[34] C. BOUVEYRON, S. GIRARD, C. SCHMID. *Une méthode de classification des données de grande dimension*, in "37èmes Journées de Statistique organisées par la Société Française de Statistique, Pau", juin 2005.

[35] C. Bouveyron, S. Girard, C. Schmid. *Une nouvelle méthode de classification pour la reconnaissance de formes*, in "20e colloque GRETSI sur le traitement du signal et des images, Louvain-la-Neuve, Belgium", 2005.

[36] J. Diebolt, M. Garrido, S. Girard, J. Ecarnot. *The EXTREMES software*, in "Fourth Conference on Extreme Value Analysis. Probabilistic and Statistical Models and their Applications, EVA 2005, Gothenburg, Suède", aout 2005.

[37] F. Forbes, M. Vignes. *Champs de Markov cachés et fusion de données individuelles et pairées pour l'identification de groupes de gènes*, in "JOBIM, Lyon, France", 2005.

[38] A. Gannoun, S. Girard, C. Guinot, J. Saracco. *Reference curves estimation via Sliced Inverse Regression*, in "International Symposium on Applied Stochastic Models and Data Analysis, ASMDA 2005, Brest", mai 2005, p. 1484–1492.

[39] L. Gardes, S. Girard. *Estimating extreme quantiles of Weibull-tail distributions*, in "Statistics for dependent data, STATDEP 2005, Paris-Malakoff", janiver 2005.

[40] L. Gardes, S. Girard. *Inférence statistique pour les lois à queue de type Weibull*, in "37èmes Journées de Statistique organisées par la Société Française de Statistique, Pau", juin 2005.

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

[42] P. Gonçalvès, H. Carrão, A. Pinheiro, M. Caetano. *Land cover classification with Support Vector Machine applied to MODIS imagery*, in "Proceedings of the 25th EARSeL Symposium, Porto (Portugal)", June 2005.

[43] J. Jacques, C. Biernacki. *Analyse discriminante généralisée : cas des données binaires avec modèles des classes latentes*, in "Premières Rencontres des Jeunes Statisticiens, Aussois", 2005.

[44] J. Jacques, C. Biernacki. *Discrimination généralisée : cas des données binaires avec modèles des classes latentes*, in "Colloque Data Mining et Apprentissage Statistique, Applications en Assurance, Niort", 2005.

[45] P. Oliveira, P. Gonçalvès, M. Caetano. *Land cover time profiles from linear mixture models applied to MODIS images*, in "Proceedings of the 31st International Symposium on Remote Sensing of Environment, St. Petersburg (Russian Federation)", June 2005.

[46] N. Peyrard, F. Forbes, D. Allard. *Comparaison de deux modélisations pour la classification de données géostatistiques*, in "37èmes Journées de Statistique organisées par la Société Française de Statistique, Pau", juin 2005.

[47] B. Schrerrer, M. Dojat, F. Forbes, C. Garbay. *Segmentation Markovienne distribuée et coopérative des tissus et des structures présents dans des IRM cérébrales*, in "RFIA, Tours, France", to appear, 2006.

Internal Reports

[48] C. Bouveyron, S. Girard, C. Schmid. *Analyse Discriminante de Haute Dimension*, Technical report, n° RR-5470, INRIA, 2005, http://www.inria.fr/rrrt/rr-5470.html.

[49] J. Diebolt, L. Gardes, S. Girard, A. Guillou. *Bias-reduced estimators of the Weibull tail-coefficient*, Technical report, n° RR-1078, LMC, 2005, http://hal.ccsd.cnrs.fr/ccsd-00008881.

[50] J. Diebolt, A. Guillou, L. Gardes, S. Girard. *Bias-reduced extreme quantile estimators of Weibull tail-distributions*, Technical report, n° RR-1080, LMC, 2005, http://hal.ccsd.cnrs.fr/ccsd-00015778.

[51] F. Forbes, G. Fort. *A convergence theorem for variational EM-like algorithms: application to image segmentation*, Technical report, n° RR-5721, Inria Rhône-Alpes, 2005, http://www.inria.fr/rrrt/rr-5721.html.

[52] L. Gardes, S. Girard. *Comparison of Weibull tail-coefficient estimators*, Technical report, n° RR-1079, LMC, 2005, http://hal.ccsd.cnrs.fr/ccsd-00009006.

[53] L. Gardes, S. Girard. *Estimation of the Weibull tail-coefficient with linear combination of upper order statistics*, Technical report, n° RR-5571, INRIA, 2005, http://www.inria.fr/rrrt/rr-5571.html.

[54] P. Gonçalvès, C. Lenoir, C. Heymes, B. Swynghedauw, C. Lavergne. *Statistical Modelling of Cardiovascular Data. An Introduction to Linear Mixed Models*, Technical report, n° RR-5787, INRIA, 2005, http://www.inria.fr/rrrt/rr-5787.html.

Miscellaneous

[55] G. Dewaele, F. Devernay, R. Horaud, F. Forbes. *The alignment between 3D-data and articulated shapes with bending surfaces*, to appear, 2006.

[56] A. Echenim. *Etude de modèles d'Analyse Discriminante de Haute Dimension*, 2005.

Bibliography in notes

[57] W. Byrne, A. Gunawardana. *Convergence theorems of Generalized Alternating Minimization Procedures*, in "Journal of Machine Learning Research", vol. 1, 2004, p. 1-48.

[58] G. Celeux, F. Forbes, N. Peyrard. *Modèle de Potts avec champ externe et algorithme de type EM pour la segmentation d'image*, in "RFIA, Toulouse, France", 2004.

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

[60] G. McLachlan, T. Krishnam. *The EM algorithm and extensions*, John Wiley, New York, 1997.

[61] R. Neal, G. Hinton. *A view of the EM algorithm that justifies incremental, sparse and other variante*, in "Learning in Graphical Models", M. Jordan (editor). , MIT Press, 1998, p. 355-368.