INRIA Research Project Proposal

MISTIS

Modelling and Inference of Complex and Structured Stochastic Systems

Thema Cognitive Systems (A): statistical modelling and machine learning

March 17, 2006

Contents

1 Short description

The team MISTIS aims at developing statistical methods for dealing with complex systems, complex models and complex 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 consider involve mixture models, Markov models, and more generally hidden structure models on one hand, and semi and non-parametric methods on the other hand. We mainly focus on two directions of research:

- Dealing with complex phenomenons. This is twofold, we have to deal with complex models and with complex data. To account for complex phenomenons, we propose to use structured models and methods allowing easy interpretations. At the models level, we propose to develop model selection and approximation techniques for complex structure models. At the data level, we study parametric models adpated to high-dimensional data and dimension reduction techniques based on non linear data analysis.

- The theoretical and practical behavior of methods. We focus on approximation justifications, asymptotic behavior and convergence analysis of our algorithms and estimators.

2 Team

Team leader

Florence Forbes [Research scientist,CR1 INRIA]

Research scientists

Paulo Gonçalves [Research scientist, CR1 INRIA, until March 2006, member of team RESO, INRIA Rhône-Alpes from April 2006]

Ph. D. students

Juliette Blanchet [MENRT, UJF, co-advised by F. Forbes and C. Schmid, team Lear, INRIA Rhône-Alpes]

Charles Bouveyron [MENRT, UJF, co-advised by S. Girard and C. Schmid, team Lear, INRIA Rhône-Alpes]

Matthieu Vignes [AC, ENS Lyon, co-advised by F. Forbes and G. Celeux, team Select, INRIA Futur]

Post-doctoral fellows

Chibiao Chen [INRIA, December 2005-December 2006]

Monica Benito [ERCIM, February 2006-October 2006]

Research scientists (partner)

Henri Bertholon [Faculty member , CNAM, Paris, 20%]

Gersende Fort [Research scientist, CR1 CNRS, TSAC/ENST, Paris, 20%]

Laurent Gardes [Faculty member, MdC UPMF, LABSAD, Grenoble, 30%]
Stéphane Girard [Faculty member, MdC UJF, HDR, SMS/LMC, Grenoble, 30%]

3 Context and overall objectives

The goal of our team is to propose statistical tools in the context of highly structured stochastic systems. Highly structured stochastic systems refer to a modern strategy for building statistical models for challenging real-world problems, for computing with them, and for interpreting the resulting inferences. The power of this approach is to handle complexity by working up from simple local assumptions in a coherent way, and that is the key to modelling, computation, inference and interpretation. Highly structured stochastic systems combine local relations to build stochastic models that exhibit great complexity. Such stochastic models have found applications in areas as diverse as signal and image processing, genetics and epidemiology. The needs of these areas have in turn stimulated important theoretical developments. Statistical methods that were once restricted to specialist statisticians, such as generalized linear modelling and multivariate discrimination, are now widely used by individual scientists, engineers, and social scientists, aided by statistical packages.

However, these powerful and flexible techniques are still restricted by necessary simplifying assumptions, such as precise measurement and independence between observations, and it long ago became clear that in many areas such assumptions can be both influential and misleading. There are several generic sources of complexity in data that require methods beyond the commonly-understood tools in mainstream statistical packages.

Often data exhibit complex dependence structures, having to do for example with repeated measurements on individual items, or natural grouping of individual observations due to the method of sampling, spatial or temporal association, family relationship, and so on.

Other sources of complexity are connected with the measurement process, such as having multiple measuring instruments or simulations generating high dimensional and heterogeneous data or such that data are dropped out or missing.

A typical example is that of an epidemiology study. The aim is to study the occurrence of a disease in a human population, especially the cause and transmission of the disease. Relevant data can be very heterogeneous. It usually includes qualitative variables such as patients answers to tests or forms and quantitative responses such as MRI or other medical analysis. Great and complex dependencies intrinsincly arise between different analysis or exams and between different individuals. In addition, data are often incomplete or missing due to patient history (death, confidentiality, etc.).

For many years, such complications in data-generating processes were often cavalierly ignored in statistical analysis, leading to unquantifiable biases. Our aim is to contribute to statistical modelling offering theoretical concepts and computational tools to handle properly some of the challenges raised by modern data. Many modern statistical modelling techniques proved capable of building highly complex probabilistic structures, to handle a great variety of practical problems.

As regards dependencies and locality, a central role is played by the concept of conditional independence. It provides a precise description of the information conveyed by the value of

one variable about others in a statistical model. Markov properties are statements about conditional independence assumptions and *Markov models* are the central subject of our research. The concept of conditional independence, whereby each variable is related locally (conditionnally) to only a few other variables, is the key to both the construction and analysis of such models.

When dealing with missing data, *mixture models* are a central starting point. They lead naturally to more general hidden structure models. Hidden structure models are also 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 exist 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 as above. 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). Expertise in this field provides us with a more general view of various problems.

Our two main domains of research are Markov models and Semi and non-parametric methods. Mixture models are a third topic of slightly less importance because it is partly included in our Markov models activity. However, its importance in applications and some specificities requires a separate presentation.

The topics we propose started individually with the former IS2 team and were running there through studies being carried out mainly independently. However half of the participants in MISTIS are new and the novelty in this proposal is the focus on the key idea of structure in models and data. This focus is unifying and various applications illustrate how we managed to merge our two main domains of expertise. Our aim is to contribute to a general formalism embracing the techniques mentioned, generalizing the ingredients of the models, broadening the scope of aplications and allowing cross-fertilization between different areas. We aim at providing a unifying framework for dealing with all the sources of complexity identified above and of many more besides.

4 Scientific foundations

A general presentation of our main domains of expertise, in this section, is followed by more specific research topics in Section 5.

4.1 Mixture models

Participants: Juliette Blanchet, Charles Bouveyron, Florence Forbes, Gersende Fort, Stéphane Girard, Matthieu Vignes.

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

Mixture models and more specifically Gaussian mixture models are among the most statistically mature methods for clustering and are also used intensively for density estimation. They correspond to a form of clustering in which we assume that individual datapoints are generated by first choosing one of a set of multivariate distributions (typically Gaussian) and then sampling from it. The success of mixture models lies partly in the fact that clustering can be seen as a labeling problem and therefore corresponds to many problems in practice.

The labeling problem. A labeling problem is specified in terms of a set of sites and a set of labels. A site often represents an item, a point or a region in the Euclidean space such as an image pixel or an image feature such as a corner, a line segment or a surface patch. A set of sites may be categorized in terms of their regularity. Sites on a lattice are considered as spatially regular (*eg.* the pixels of a 2D image). Sites which do not present spatial regularity are considered as irregular. This is the usual case when sites represent geographic locations, features extracted from images at a more abstract level, such as the detected corners and lines and more generally *interest points*. It can also be that the sites correspond to items (*eg.* genes) that are related to each other through a distance or dissimilarity measure or simply to a collection of independent items.

A label is an event that may happen to a site. We will consider only discrete label set. In this case, a label assumes a discrete value in a set of K labels. In edge detection, for example, the label set is the two component set $\{edge, non-edge\}$.

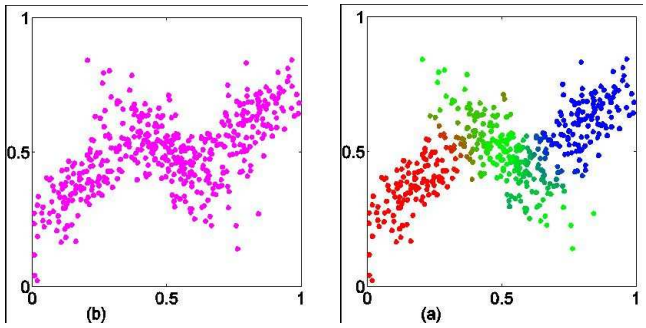
The labeling problem is to assign a label from a label set $\mathcal{L}$ to each of the sites. If there are n sites, the set $z = \{z_1, \ldots, z_n\}$ with $z_i \in \mathcal{L}$ for all $i \in S$, is called a labeling of the sites in S in terms of the labels in $\mathcal{L}$. When each site is assigned a unique label, a labeling can be regarded as a function with domain S and image $\mathcal{L}$. In mathematical programming a labeling is also called a coloring, in the terminology of random fields it is called a configuration. In vision, it can corresponds to an edge map, an interpretation of image features in terms of object features, or a pose transformation and so on.

Our approach of the labeling problem is based on mixture models and more generally on hidden structure models. 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 $y = y_1, \ldots, y_n$ and unobserved or missing data $z = z_1, \ldots, z_n$. The missing data z_i represents the memberships to one of a set of K alternative categories, *ie.* the labels. The distribution of an observed y_i can be written as a finite mixture of distributions,

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

Besides their usefulness for clustering, mixture models are a very flexible method of modelling. As any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities, mixture models provide a convenient semi-parametric framework in which to model unknown distributional shapes, whatever the objective.

These models are also 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*



(a) (b)

Figure 1: EM Clustering: (a) data with unknown class assignments (missing data), (b) membership probabilities using EM.

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 [6,8] in missing data problems. The algorithm iteratively maximizes the expected complete-data log-likelihood over values of the unknown parameters, conditional on the observed data and the current values of those parameters. In the clustering context, it provides parameters estimation but also values for missing data by providing membership probabilities to each group (Figure 1).

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 and learning (Section 5.1.1).

When the z_i's are not independent, the inter-relationship between sites can be maintained by a so-called neighborhood system usually defined through a graph. In this case, we will rather consider Hidden Markov Models as presented in the following section. In both cases, the high-dimensionality of the observations (y_i's) may be a problem in practice and we also investigated techniques of dimension reduction in a classification context (Sections 5.1.2 and 5.2.1).

4.2 Markov models

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

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

Graphical modelling provides a diagrammatic representation of the logical structure of a

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

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

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 (Section 5.2.1). This also requires a good understanding of the role of each parameter and methods to tune them depending on the goal in mind (Section 5.2.2). As regards, estimation algorithms, they correspond to an energy minimization problem which is NP-hard and usually performed through approximation. We focus (Section 5.2.4) 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 (Section 5.2.5). We also propose some tools for model selection (Section 5.2.6). Eventually we investigate ways to extend the standard Hidden Markov Field model to increase its modelling power (Section 5.2.3).

4.3 Semi and non parametric methods

Participants: Charles Bouveyron, Laurent Gardes, Stéphane Girard, Paulo Gonçalves.
Keywords: non parametric methods, boundary estimation, multiresolution analysis, wavelets.

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* (Section 5.3.1) and *image segmentation* (Section 6.1.5). These techniques are also of great use for dimension reduction since they allow to avoid assumptions on the observations distribution (Section 5.3.2). 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 (Section 5.3.3).

5 Domains of research

5.1 Mixture models

In the last few years, mixture models became a widely used approach to statistical pattern recognition. Efficient applications relie on the development of models and techniques allowing learning of the patterns of interest at a reasonable cost.

5.1.1 Learning and classification techniques

Participants: Juliette Blanchet, Charles Bouveyron, Florence Forbes, Gersende Fort, Stéphane Girard, Matthieu Vignes.

The objective is to develop classification algorithms which will be tested and used within specific contexts (*e.g* object recognition and texture description, see Section 6). We will base our approach on mixture models (possibly with missing data and spatial constraints) and the EM algorithm. In particular, we will seek for a practical solution to the problem of speed associated with EM, which is sometimes too slow for the kind of applications to be addressed here. We already proposed and studied the so-called CEMM algorithm [1] for this purpose and will go on investigating along this line.

Unsupervised learning. This field includes a lot of ad-hoc and not unified methods. Our approach is based on mixture models and more generally on hidden structure models such as hidden Markov models which can account for additional spatial constraints (see Section 5.2). When the hidden data is large and the hidden structure complex, for instance in image analysis there are often as many hidden variables as observed ones, estimation algorithms for the models parameters are far from being suitably based. Our aim is to investigate their theoretical foundations and to optimize their use.

Semi supervised or partially supervised learning. In partially supervised learning, only part of the data is labelled, typically few positive examples. It is often the case in computer vision since unlabelled data are easy to obtain while it can be very costly to label data specifically. Therefore we will consider an intermediate approach between classification (un-

supervised learning) and discrimination (supervised learning) based on [2,1]. This approach is based on hierarchical mixture models in the spirit of "Mixture Discriminant Analysis" from [11]. Each class is modelled as a mixture model and each point in the learning set can be labelled (discrimination problem), unlabelled (classification problem) or partially labelled. A partial label means that an observation is for instance known not to be in class C but can be both in class A or B. If the learning set is only made of labelled and unlabelled data, we refer to the problem as semi-supervised. The associated algorithm is derived from the EM algorithm.

Overlapping classes. Usual classification techniques assign each object to a single class. However, in a variety of important applications, overlapping clustering, wherein some items are allowed to be members of two or more discovered clusters, is more appropriate. For instance, in biology, genes have more than one function by coding for proteins that participate in multiple metabolic pathways. Another example is that of movies databases where many movies belong to more than one genre such as action, science fiction genres, etc. Methods have been proposed [3] that are based on mixtures, providing empirical evidence that accounting for overlapping clusters within a model produces more accurate clusters than an alternative naive method based on thresholding the membership probabilities in a traditional soft clustering. We will also make the connection with the partial volume problem in medical imaging (MRI). The limited spatial resolution of MR imaging and the complex shape of the tissues interfaces imply that a large part of the voxels in MR images are so-called partial volume voxels, *ie.*, voxels that contain not a single tissue but rather a mixture of two or more tissue types. We will investigate how the EM approach can be extended to take the partial volume effect into account. Starting from ideas in [21], we will investigate how it can be extended outside the MRI application (*eg.* genetic data analysis) and how it then compares to the approach in [3].

5.1.2 Taking into account the curse of dimensionality.

Participants: Juliette Blanchet, Charles Bouveyron, Florence Forbes, Laurent Gardes, Stéphane Girard, Paulo Gonçalves, Monica Benito (ERCIM Post-doc).

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), we propose new

[2] C. Ambroise and G. Govaert. EM algorithm for partially known labels. In *Data Analysis, Classification, and Related Methods, Springer. Proc. 7th Conf. Int. Federation of Classication Societies (IFCS-2000)*, 2000.

[1] C. Ambroise, T. Denoeux, G. Govaert, and P. Smets. Learning from an imprecise teacher : probabilistic and evidential approaches. In *Proceeding of ASMDA 2001*, Compiègne, France, 2001.

[11] T Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *J. Royal. Statist. Soc. B.*, 58:155–176, 1996.

[3] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. Mooney. Model-based overlapping clustering. In *Proc. 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 532–537, Chicago, United States, August 2005.

[21] K. Van Leemput, D. Vandermeulen, and P. Suetens. Unifying framework for partial volume segmentation of brain MR Images. *IEEE Trans. on Medical Imaging*, 22(1):105–119, 2003.

Gaussian models of high dimensional data for classification purposes [2, 3]. 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 [4, 5]. Some versions of this method have been tested on the supervised classification of objects in images (Section 6.1.1). We are currently working on the adaptation of this approach, named HDDC for High Dimensional Data Clustering, to the unsupervised classification framework (see also Section 6.1.1). 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 (Section 5.2.1) and obtained significant improvement in applications such as texture recognition where the observations are high-dimensional (Section 6.1.2). 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 Section 6.1.5).

We are then also willing to get rid of the Gaussian assumption. To this end, non linear models and semi parametric methods are necessary (see Section 5.3.2).

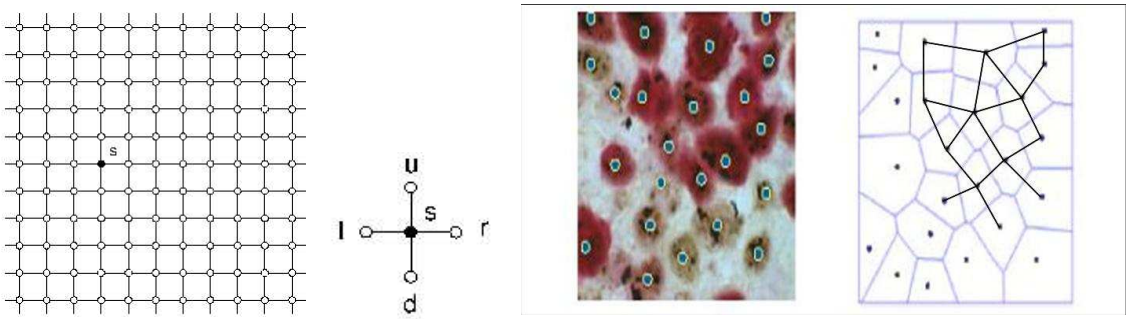
5.2 Markov models

The first three following sections deal explicitly with Markov models for sites or items at irregular locations. This goes beyond the standard regular lattice case and requires some adaptation. It implies additional isues such as the choice of the neighborhood structure which may depend on the application (see Figure 2). Indeed, for irregular lattices, the points relative displacements do not follow a predictable pattern and their linkage are not always obvious from their geometry so that a lot of possible spatial structures can be generated. In Section 5.2.3, we investigate how our expertise in regular Markov models could transpose to and profit from techniques used in geostatistics. The following sections apply for all type of structure but the regularity of the sites is not specifically under consideration. They concern the theoretical (convergence) properties of our algorithms and the question of how to select the best or more appropriate models.

5.2.1 Markov models for the spatial organization of image descriptors.

Participants: Juliette Blanchet, Florence Forbes.

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



(a) (b) (c)

Figure 2: Structures: (a) regular lattice and (b) a standard four nearest neighbor system, (c) biological image of cells and proposed structure based on Voronoi tesselation.

regions. For tractability, most approach 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-occurence 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 choose 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 strenght 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 (Section 5.2.4) and to investigate how to make them accurate and computationnaly 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 (see Section 5.1.2) we show that recognition rates can be improved and that promising results can be obtained using a general statistical formalism. We focus in particular on texture recognition (see Section 6.1.2) but further work includes other contexts such as object recognition and tracking. This is the basis of the PhD work of Juliette Blanchet (co-advised by Cordelia Schmid from team LEAR).

5.2.2 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.

As regards classification issues, 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 distance or some similarity measure defined on the objects. In practice such data can be missing and EM offer a good framework to deal with this case (see [8]). 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 are weighted according to pair data reflecting 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. An illustration is given in Section 6.2.1 with an application to genetic data analysis.

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 [6], we investigate the role of *singleton potentials* which are parameters usually ignored in standard Markov model-based segmentation. In [7] we use 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 Section 6.1.6 for an application to MRI analysis). This last work is interesting in that it illustrates the successful use of our statistical methodology to solve a difficult imagery issue. Besides, we aim at generalizing the approach by focusing on the issue of modelling and taking into account *high-level* a priori knowledge in the context of a new European project POP (see Section 10.1). In addition to further investigating the use of integrated models in our collaboration with INSERM/TIMC (see Section 6.1.6), we hope to generalize the approach to the modelling of visual and auditory perception. A PhD student will be hired in June 2006 to investigate these aspects in POP.

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

---

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

5.2.3 Comparing two models for clustering spatial data

Participant: Florence Forbes.

This is joint work with Denis Allard and Nathalie Peyrard from INRA, Avignon.

Our expertise in Markov models includes the investigation of other related modelling solutions that offer similar processing advantages and possibly superior modelling capabilities. The so-called pairwise and triplets Markov random fields ([4]) go in this direction and can be related to the extensions we consider. With this in mind we focus on the problem of clustering spatial data *ie.* of grouping observations measured at different points of the plane, taking into account the geographical distances between the points. We compare two models to solve this problem: hidden Markov random fields (HMRF) and a geostatistical model which uses Gaussian random fields. Geostatistical-type problems are distinguished most clearly from lattice-type problems by the ability of the spatial index to vary continuously. We plan to study how methods from one class of problems can be borrowed from methods usually associated with another class. As regards parameter estimation we plan to investigate an EM-like algorithm. If approximations of this algorithm exist for HMRFs (see Section 5.2.4), this is not the case for geostatistical models. We propose [8] an approach mimicing the solution in the independant case. Models performance is compared on simulated and real data: they show the superiority of the geostatistical model in terms of classification error but the classifications obtained with the HMRF model are smoother and sometimes more satisfactory. Further work is required to better understand the possible extensions of standard HMRF.

5.2.4 EM procedures using mean field-like approximations for Markov model-based image segmentation and clustering

Participants: Juliette Blanchet, Florence Forbes, Gersende Fort, Chibiao Chen (INRIA post-doc).

The EM algorithm mentionned in Section 4.1 has an elegant formulation and when it is applied to appropriate model structures it yields parameter update procedures that are easy to derive and straightforward to implement. However, outside simple or standard cases, the EM algorithm yields update procedures that do not have closed form expressions and it is seldom tractable analytically. In particular, when focusing on image segmentation and Markov Random Fields (MRF) estimation, difficulties arise due to the dependence structure in the models and approximations are required. A heuristic solution using mean field approximation principle has been proposed in [22]. The mean field approach consists of calculating quantities related to a complex probability distribution, by using a simple tractable model such as the family of independent distributions. Using ideas from this principle, we propose [9] in the context of Markovian image segmentation a class of EM-like algorithms generalizing [22] which show good performance in practice.

[4] D. Benboudjema and W. Pieczynski. Unsupervised image segmentation using triplet markov fields. *Computer Vision and Image Understanding*, 99(3):476–498, 2005.

[22] J. Zhang. The Mean Field Theory in EM Procedures for Markov Random Fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583, 1992.

The idea underlying these algorithms is to replace the intractable Markov distribution by a simpler distribution obtained by fixing the neighbors of each pixel to constant values. Then, an iteration of a mean field-like algorithm consists of two steps: in the first step the values for the neighbors are updated according to the observations and to the current value of the parameter. It follows an approximation of the intractable Markov distribution. The second step consists of carrying out the EM algorithm for the corresponding approximated observed likelihood to obtain an updated value of the parameter. Mean field-like algorithms can thus be related to the EM algorithm for independent mixture models, with the significant difference that the mixture model adaptively changes at each iteration depending on the current choice of the neighbors values. In [9], we compare three different ways of updating the neighbors in the first step: the mean field approximation of the conditional mean (*mean field algorithm*), an approximation of the conditional mode (*mode field algorithm*) and a simulated realization of the conditional Gibbs distribution obtained with the Gibbs sampler of [10] (*simulated field algorithm*). The last choice led to a new stochastic algorithm which appeared to be the most promising for its good performance on synthetic and real image experiments. As mentionned in many sections of the present document, we are working on extending the modelling capabilities of these kind of algorithms and on providing a better understanding of their properties. See for instance, Section 6.2.2 for an on going investigation on non image data.

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

Participants: Florence Forbes, Gersende Fort.

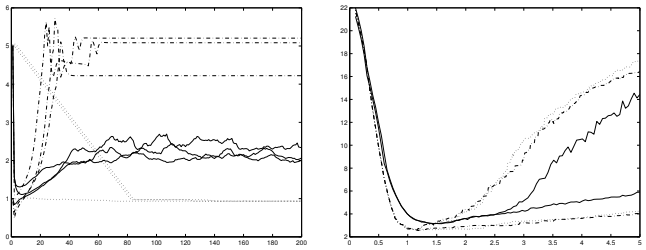
For the standard EM algorithm, parameter estimates yield increasing likelihood over the observed data and the convergence behavior of this process is well understood. However, since it is often the case that there are no other feasible choices rather than to resort to the mean field approximation in 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 [5] and [17] but these variants cannot be applied in the MRF segmentation framework and further approximations are required. We are investigating [10] 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 [5] that it can be applied to perform image segmentation tasks and so on the basis of theoretical convergence results.

---

[10] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[5] W. Byrne and A. Gunawardana. Convergence theorems of Generalized Alternating Minimization Procedures. *Journal of Machine Learning Research*, 1:1–48, 2004.

[17] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variante. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, 1998.



(a) (b)

Figure 3: Estimation and effect of the spatial parameter β in standard Hidden Potts model (HMRF): (a) β estimates trajectory versus the number of iterations for different parameter starting values, with *Mean Field* (dot line), *Simulated Field* (solid line) and *MCVEM* (dash-dot line), (b) Classification error rate versus β obtained by *Mean Field* (dot line), *Simulated Field* (solid line) and *MVCEM* (dash-dot line), when the segmentation algorithm is started from two different initial classifications.

The basis of our work is the paper [11] which focus 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 [5]. Experiments on synthetic and real images shows that the algorithm performance is very closed and sometimes better to that of [9]. 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 (see Figure 3 for an illustration of practical implementation issues). In particular, analysing how simulation step should be incorporated so as to interact advantageously with deterministic approximations seems promising.

5.2.6 Approximations for selecting complex structure models

Participant: Florence Forbes, Gersende Fort.

Choosing the probabilistic model that best accounts for the observations is an important first step for the quality of the subsequent statistical inference and analysis. In most cases the choice is done subjectively using expert knowledge or *ad hoc* procedures and there is a striking lack of systematic data-based approaches. When recasting this choice as a problem of probabilistic model comparison, most selection criteria involve calculating integrated likelihoods for a number of models, *i.e.* the likelihoods of the observations integrated over the respective model parameters. For a lot of models of interest, these integrated likelihoods are high dimensional and intractable integrals so that most available software is generally inefficient for their evaluation. Various approximations have been proposed. In particular the Bayesian Information Criterion (BIC) approximation of [19] is based on the Laplace method

[5] W. Byrne and A. Gunawardana. Convergence theorems of Generalized Alternating Minimization Procedures. *Journal of Machine Learning Research*, 1:1–48, 2004.

[19] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

for integrals. Many other approaches can be found in the literature on model selection (see for instance the list of references in [14]). BIC has become quite popular due to its simplicity and its good results in cases where other standard model selection procedures were unsatisfactory.

In the context of Markov model-based image segmentation, we propose variational approximation tools [13] to deal with this issue in practice. Our aim is to apply these techniques to built and select models of interest, possibly models with a complex structure. In these situations, exact calculation is not possible and simulation methods such as Monte Carlo Markov Chains (MCMC) methods reach their limits. In [12], we focus on the use of BIC for the underlying issue of choosing a model from a collection of hidden Markov random fields. In this case, we have no specific results on the quality of BIC as an approximation of the integrated likelihood and this choice as a selection criterion is arguable. However, the question of the criterion ability to asymptotically choose the correct model can be addressed independently of the integrated likelihood approximation issue. As an illustration, the author in [9] proved recently that for the more specialized but related case of hidden Markov chains, under reasonable conditions, the *maximum penalized marginal likelihood* estimator of the number of hidden states in the chain is consistent. This estimator is defined for a class of penalization terms that includes the BIC correction term and involves an approximation of the maximized log-likelihood which is not necessarily good, namely the maximized log-marginal likelihood. In particular, this criterion is consistent even if there is no guarantee that it provides a good approximation of the integrated likelihood. This suggests that a good approximation of the maximized log-likelihood is not a strong requirement to obtain consistent criteria. A key point in [9] seems to be the decomposition of the criterion as a sum of identically distributed terms. The criteria proposed in this paper can also be written as sum because of the factorization property of the distributions involved. The generalization is not straightforward but our next step is therefore to investigate if consistency results can be deduced in a similar way.

At a different level, we also investigated the deviance information criterion (DIC) introduced by Spiegelhater et al. [20]. It is directly inspired by linear and generalized linear models, but it is not so naturally defined for missing data models. We have reassessed the criterion for such models, testing the behavior of various extensions in the cases of independent mixture and random effect models. This is joint work with G. Celeux from team SELECT, INRIA futur, Mike Titterington (Univ. of Glasgow, Scotland) and Christian Robert, (CEREMADE, Paris Dauphine) [13]. As illustrated by the associated discussion in [13], a lot of questions and issues are still open regarding the use of DIC as a measure of complexity and a well-grounded criterion.

---

[14] R. Kass and A. Raftery. Bayes factor. *Journal of the American Statisticial Association*, 90:733–795, 1995.

[13] M.I. Jordan, editor. *An introduction to variational methods for graphical models.* MIT Press, 1999.

[9] E. Gassiat. Likelihood ratio inequalities with application to various mixtures. *Annales de l'institut Poincaré*, 2002.

[20] D. J. Spiegelhalter and al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, series B*, 64:1–34, 2002.

5.3 Semi and non parametric methods

We have expertise in techniques that can be referred to as semi and non-parametric methods. They include projection methods that decompose signals and images on a set of functions such as wavelets (Sections 5.3.1 and 5.3.3) and kernel methods (Section 5.3.1).

5.3.1 Boundary estimation

Participants: Laurent Gardes, Stéphane Girard.

This is joint work with Anatoli Iouditski (Univ. Joseph Fourier, Grenoble), Guillaume Bouchard (Xerox), Pierre Jacob and Ludovic Menneteau (Univ. Montpellier 2) 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 (see Figure 10 in Section 6.2.3). 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.

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 propose estimators based on projection as well as on kernel regression methods applied on the extreme values set [14], for particular set of points. In this specific framework, we can obtain the asymptotic distribution of the estimators. In his PhD work, co-directed by Pierre Jacob and Stéphane Girard, Laurent Gardes [15] 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.

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 [16]. The resulting estimate is parsimonious, it only relies on a small number of points (see Figure 4). This method belongs to the Support Vector Machines (SVM) techniques. Their finite sample performances are very impressive but their asymptotic properties are not very well known, the difficulty being that there is no explicit formula of the estimator. However, such properties are of great interest, in particular to reduce the estimator bias. Two directions of research will be investigated. The first one consists in modifying the optimization problem itself. The second one is to use *Jacknife* like methods, combining two biased estimators so that the two bias cancel out. One of the goals of our work is also to establish the speed of convergence of such methods in order to try to improve them.

See Section 6.2.3 for an application to real data.

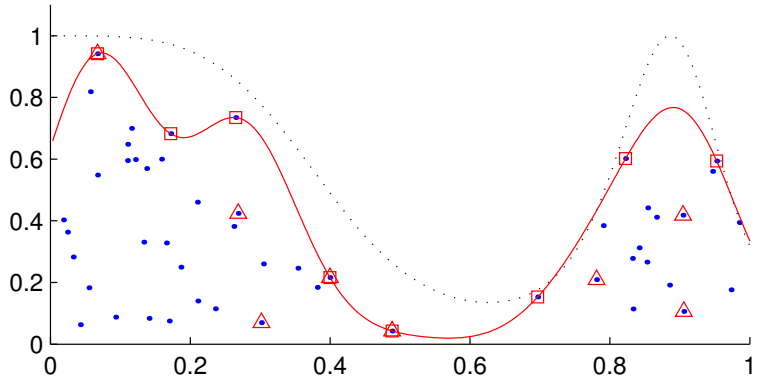


Figure 4: Boundary estimation using a linear programming approach: blue points are data points, the dot curve is the unknown boundary, the continuous curve is the estimated boundary, the square points are points on the estimated boundary and the triangle points are the additional points actually used in the estimation (support vectors).

5.3.2 Non parametric view of high dimensional data

Participants: Charles Bouveyron, Stéphane Girard.

As mentionned in Section 5.1.2, our work on high dimensional data includes non parametric aspects. They are related to Principal Component Analysis (PCA) which is traditionnaly used 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 (see Figure 5). When dealing with classification problems, such as in Section 5.1.2, our main project is then to adapt the non linear PCA method proposed in [17]. 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 [18].

5.3.3 Multiresolution Analysis and Markov tree models

Participants: Florence Forbes, Paulo Gonçalves.

This is joint work with Jean-Baptiste Durand from LMC, Grenoble.

For a large class of signals and images, orthogonal wavelet decompositions provide a parsimonious representation where only few coefficients have a significant non-zero amplitude. Moreover, although stricto sensu they do not correspond to Kharunen-Loéve basis, in many cases it is reasonable to neglect residual correlations between wavelet coefficients. In a former work with J. B. Durand (LMC, Grenoble), we investigated the opposite situation where it is important to take into account these correlations [19]. This is notably the case for scaling law processes (such as fractional Brownian fields) where dominant correlation is inter-scale, and for spatially structured images where the intra-scale correlation prevails. We proposed to model these interactions by means of hidden Markov trees applied to the dyadic structure of multiresolution decompositions. The resulting statistical model underlies a hidden state

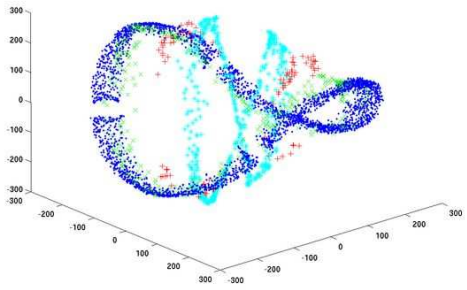


Figure 5: Highly non linear data: 128-dimensional SIFT descriptors extracted from successive rotations of a single image and projected in $I\!R^3$ for visualization. The colors correspond to different scales.

tree associated to the wavelet coefficients tree, and we proposed efficient forward-backward type algorithms for likelihood maximization.

Hence, if the time axis (for time series) or the spatial axes (for images) are at aim, we strive at modelling the statistical dependence of a system state conditionally to its past (for time series) or to its spatial context (for images). Conversely, if one focuses on the scale axis, the model will emphasize the multi-scale interactions in time (for time series), respectively in space (for images). An application of this model is given in Section 6.1.4.

6 Application domains

6.1 Image Analysis

Participants: Juliette Blanchet, Charles Bouveyron, 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 at INRIA Rhône-Alpes, issues about object and class recognition and about the extraction of visual information from large image data bases (Sections 6.1.1 and 6.1.2).

Other applications in medical imaging are natural. We work more specifically on MRI data in collaboration with INSERM and TimC (Section 6.1.6) and with the Statistics Department of University of Washington, Seattle (Section 6.1.7).

We also consider other statistical 2D fields coming from other domains such as remote sensing, in collaboration with Laboratoire de Planétologie de Grenoble (Section 6.1.3), and other collaborations (Sections 6.1.4 and 6.1.5).

6.1.1 Supervised and unsupervised classification of objects in images

Participants: Charles Bouveyron, Stéphane Girard.

This is joint work with Cordelia Schmid, LEAR, INRIA Rhône-Alpes.

Supervised framework. In this framework, small scale-invariant regions are detected on a learning image set and they are then characterized by the local descriptor SIFT [16]. The object is recognized in a test image if a sufficient number of matches with the learning set is found. The recognition step is done using supervised classification methods. Frequently used methods are Linear Discriminant Analysis (LDA) and, more recently, kernel methods (SVM) [12, chap. 12]. In our approach, the object is represented as a set of object parts. As an example for a motorbike, we will consider three parts: wheels, seat and handlebars.

Obtained results showed that the HDDA method described in Section 5.1.2 gives better recognition results than SVM and other generative methods. In particular, the classification errors are significantly lower for HDDA compared to SVM. In addition, HDDA method is as fast as standard discriminant analysis (computation time $\simeq$ 1 sec. for 1000 descriptors) and much faster than SVM ($\simeq$ 7 sec.).

Unsupervised framework. Our approach learns automatically discriminant object parts and then identifies local descriptors belonging to the object. It first extracts a set of scale-invariant descriptors and then learns a set of discriminative object parts based on a set of positive and negative images. Learning is "weakly supervised" since objects are not segmented in the positive images. Recognition matches descriptors of a unknown image to the discriminative object parts.

Object localization is a challenging problem since it requires a very precise classification of descriptors. For this, it is necessary to identify the descriptors of an image which have a high probability to belong to the object. The adaptation of HDDA to the unsupervised framework, called HDDC, allows to compute the posterior probability for each interest point that it belongs to the object. Finally, the object can be located in a test image by considering the points with the highest probabilities. In practice, 5 or 10 percents of all detected interest points are enough to locate efficiently the object. See an illustration in Figure 6.

We also consider the application of image classification. This step decides if the object is present in the image, i.e. it classifies the image as positive (containing the object) or negative (not containing the object). We use our decision rule to assign a posterior probability to each descriptor and each cluster. We then decide based on these probabilities if a test image contains the object or not. Previous approaches [7] have used a simple empirical technique to classify a test image. We introduce a probabilistic technique which uses the posterior probabilities. We obtain for a test image I a score S$\in [0, 1]$ that I contains the object. We decide that a test image contains the object if the score S is larger than a given threshold. This probabilistic decision has the advantage of not introducing an additional parameter and of using the posterior probability to reject (assign a low weight) to dubious points.

---

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, NewYork, 2001.

[7] Gyuri Dorkó and Cordelia Schmid. Object class recognition using discriminative local features. Submitted to IEEE Trans. on Pattern Analysis and Machine Intelligence, updated 13 September, 2005.

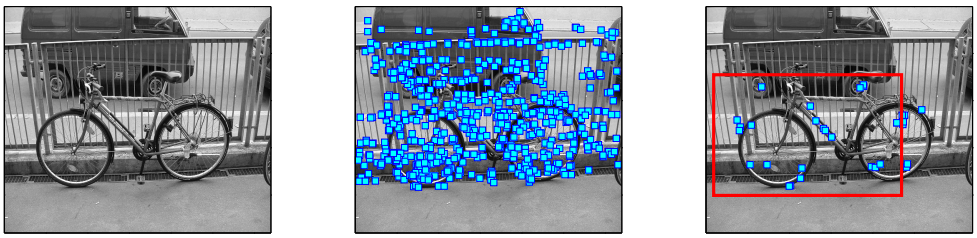
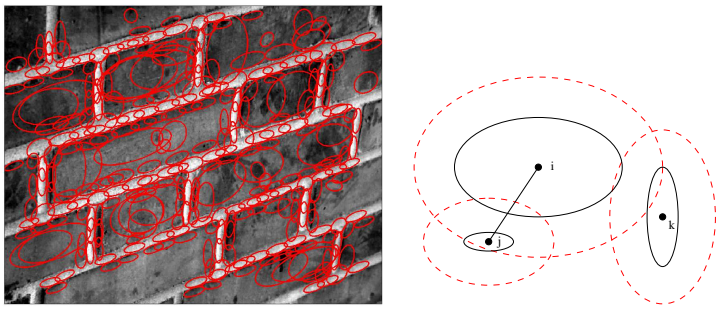


Figure 6: Object localization: from left to right, Original image, All detections, Final localization.



(a) (b)

Figure 7: Spatial relationships between descriptors: (a) detected interest points (ellipse centers) and associated scales (ellipse sizes and orientations) using Laplace detector and SIFT descriptors, (b) construction of a neighborhood system between the detected points: points within the range of an enlarged ellipse are connected to the ellipse center.

6.1.2 Markov Random Fields for recognizing textures

Participants: Florence Forbes, Juliette Blanchet.

This is joint work with Cordelia Schmid, LEAR, INRIA Rhône-Alpes.

We present a new probabilistic framework for recognizing textures in images. Images are described by local affine-invariant descriptors and by spatial relationships between these descriptors. See an illustration in Figure 7. We propose to introduce in texture recognition the use of statistical parametric models of the dependence between descriptors. We choose Hidden Markov Models (HMM) and follow the method described in Section 5.2.1. 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 very promising results [20] (see an illustration in Figure 8).

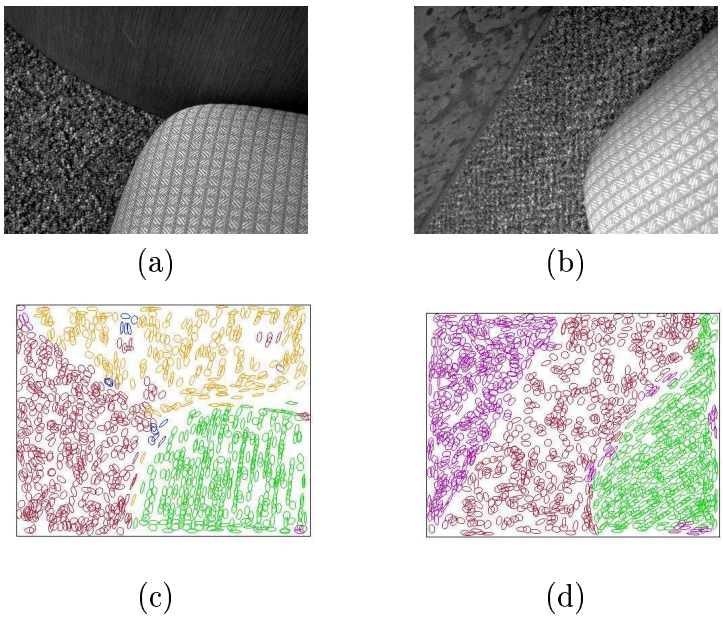


Figure 8: Texture recognition using Hidden Markov Models: (a) and (b) original multi-texture images, (c) and (d) classification results.

6.1.3 Statistical methods for the visualization and analysis of complex remote sensing data

Participants:

Monica Benito (ERCIM Post-doc), Juliette Blanchet, Florence Forbes, Laurent Gardes, Stéphane Girard.

This is joint work with Sylvain Douté and Etienne Deforas from Laboratoire de Planétologie de Grenoble, France.

Visible and near infrared imaging spectroscopy is one of the key techniques to detect, to map and to characterize mineral and volatile (eg. water-ice) species existing at the surface of the planets. Indeed the chemical composition, granularity, texture, physical state, etc. of the materials determine the existence and morphology of the absorption bands. The resulting spectra contain therefore very useful information. Current imaging spectrometers provide data organized as three dimensional hyperspectral images: two spatial dimensions and one spectral dimension.

Since 2005, a new generation of imaging spectrometers is emerging with an additional angular dimension. The surface of the planets will now be observed from different view points on the satellite trajectory, corresponding to about ten different angles, instead of only one corresponding usually to the vertical (0 degree angle) view point. Multi-angle imaging spectrometers present several advantages: the influence of the atmosphere on the signal can be better identified and separated from the surface signal on focus, the shape and size of the surface components and the surfaces granularity can be better characterized.

However, this new generation of spectrometers also results in a significant increase in the

size (several tera-bits expected) and complexity of the generated data. Consequently, HMA (Hyperspectral Multi Angular) data induce data manipulation and visualization problems due to its size and its 4 dimensionality.

We propose to investigate the use of statistical techniques to deal with these generic sources of complexity in data. This requires methods beyond the commonly-understood tools in mainstream statistical packages. Our goal is twofold:

- we will first focus on developing or adapting dimension reduction methods, classification and segmentation methods for informative, useful visualization and representation of the data previous to its subsequent analysis. We plan to combine mixture models with dimension reduction techniques ([12]) .
- We will also address the problem of physical model inversion which is important to understand the complex underlying physics of the HMA signal formation. The models taking into account the angular dimension result in more complex treatments. We will investigate the use of semiparametric dimension reduction methods such as SIR (Sliced Inverse Regression, [15]) to perform model inversion, in a reasonable computing time, when the number of input observations increases considerably.

6.1.4 Image fusion using Multiresolution Analysis and Markov tree models

Participant: Paulo Gonçalves.

This is joint work with Jean-Baptiste Durand (LMC, Grenoble), Hugo Carrão 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 [19] (see Section 5.3.3). 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

---

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, NewYork, 2001.

[15] K.C. LI. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–342, 1991.

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.

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

Participants: Charles Bouveyron, Stéphane Girard, Paulo Gonçalves.

This is joint work with Hugo Carrão and 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 Sections 5.1.2 and 5.3.2).

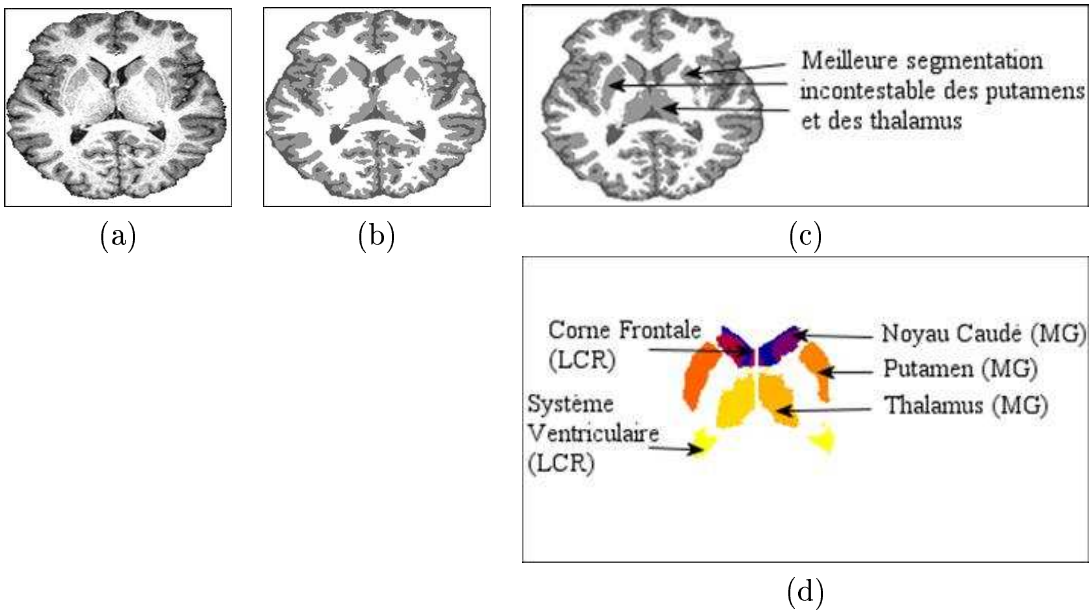
Eventually, our model parameters are used as inputs of a supervised SVM classifier. Performance is then exhaustively compared to that obtained when the same classifier is directly applied to a single date multi-spectral reflectance data. First results are reported in [21, 22].

6.1.6 Distributed and cooperative Markovian segmentation of both tissues and structures in brain MRI

Participant: Florence Forbes.

This is joint work with Benoit Scherrer, Michel Dojat and Christine Garbay from TIMC and INSERM.

Accurate tissue and structure segmentation of MRI brain scan is critical for several applications. Markov random fields are commonly used for tissue segmentation to take into account spatial dependencies between voxels, hence acting as a labelling regularization. However, such a task requires the estimation of the model parameters (eg. Potts model) which is not tractable without approximations. The algorithms in [9] (see Section 5.2.4) based on EM and variational approximations are considered. They show interesting results for tissue segmentation but are not sufficient for structure segmentation without introducing a priori anatomical knowledge. In most approaches, structure segmentation is performed after tissue segmentation. We suggest considering them as combined processes that cooperate. Brain anatomy is described by fuzzy spatial relations between structures that express general relative distances, orientations or symmetries. This knowledge is incorporated into a 2-class



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

Figure 9: Distributed and cooperative Markovian segmentation: (a) real 3T scan, (b) tissue segmentation without anatomical knowledge, (c) and (d) tissue and structure segmentations using the distributed and cooperative approach.

Markov model via an external field. This model is used for structure segmentation. The resulting structure information is then incorporated in turn into a 3 to 5-class Markov model for tissue segmentation via another specific external field. Tissue and structure segmentations thus appear as dynamical and cooperative MRF procedures whose performance increases gradually. This approach is implemented into a multi-agent framework, where autonomous entities, distributed into the image, estimate local Markov fields and cooperate to ensure consistency. 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 (See Figure 9).

6.1.7 Model-based Region-of-Interest Selection in dynamic breast MRI

Participant: Florence Forbes.

This is joint work with Chris Fraley and Adrian Raftery from University of Washington, Seattle, with Dave Goldhaber and Dianne Georgian-Smith, M.D. from Harvard Medical School, Massachussets General Hospital.

Magnetic Resonance Imaging (MRI) is emerging as a powerful tool for the diagnosis of breast abnormalities. Dynamic analysis of the temporal pattern of contrast uptake has been applied in differential diagnosis of benign and malignant lesions to improve specificity. Signal intensity time course data are useful for differentiating benign from malignant enhancing lesions. The overall shape of the time-signal intensity curve is an important criterion, while a single attribute of the curve, such as the enhancement rate, may not be enough.

Selecting a region of interest (ROI) is an almost universal step in the process of examining the contrast uptake characteristics of a breast lesion. We propose [23, 24, 25] an ROI selection method that combines model-based clustering of the pixels with Bayesian morphology [26], a more recent statistical image segmentation method. We then investigate tools for subsequent analysis of signal intensity time course data in the selected region.

Results on a data base of 19 patients are promising. The method provides informative segmentations and good detection rates are obtained. The investgation indicates that our proposed statistical methods, which enable us to take into account more than a single enhancement measure, are quite promising for tumor identification. There is a clear gain in combining segmentation with kinetics analysis. Associating the location and shape of a lesion with its pattern of uptake proved to be useful in resolving questionable cases. the trade-off between smoothness and resolution needs to be assessed by further empirical research on other images. Our study is limited to the determination of feasibility for the proposed computational methods. Clinical value would have to be assessed in more extensive and controlled studies, which in the light of our initial experience may be warranted.

6.2 Biology and Medicine

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.

6.2.1 Integrated Markov models on irregular grids for clustering gene expression data

Participants: Florence Forbes, Matthieu Vignes.

Because of the increasing amount of genetic data generated by researchers, there is a great need to develop methodologies to analyse and to use the information contained in this data. A major challenge in bioinformatics is to reveal interactions between components of living organisms and discover the corresponding networks responsible for their biological complexity. In this framework, clustering of genes into groups sharing common characteristics is a useful exploratory technique. It is frequently used as the basis for further computational analysis. As an example, the function of a gene can be predicted according to known functions of other genes from the same cluster.

A wide range of clustering algorithms have been proposed to analyze gene expression data but most of them consider the genes as independent entities or include relevant information on gene interactions a posteriori. We propose a probabilistic model that has the advantage to take into account individual features (e.g. expression) and pairwise data (e.g interaction information coming from biological networks) simultaneously. As mentionned in Section 5.2.2, our model is based on hidden Markov random fields in which parametric probability distributions account for the distribution of individual data for each gene. Data on pairs are included through a graph where the nodes represent the genes and the edges are weighted according to pair data, for instance in order to reflect distances or similarity measures between genes. This model has many interesting features. It leads to various possible

statistical criteria to select automatically the number of clusters. It is also able to incorporate many types of data. It is flexible in the sense that its generalization to include missing data, that often occur when dealing with expression data, is straightforward. Its extension to overlapping clustering methods, to deal with more realistic situations where genes can belong to many groups at the same time, can also be considered. Preliminary investigations are reported in [27]. We start illustrating and validating the approach on simulated data as well as on yeast expression data combined with pathways neighbourhoods.

6.2.2 Modelling and inference of population structure from genetic and spatial data

Participant: Chibiao Chen (INRIA post-doc), Florence Forbes.

This is joint work with Olivier François and David Robelin from team TimB in TIMC laboratory.

In applications of population genetics, it is often useful to classify individuals in a sample into populations which become then the units of interest. However, the definition of populations is typically subjective, based, for example, on linguistic, cultural, or physical characters as well as the geographic location of sampled individuals. Recently, Pritchard et al [18], proposed a Bayesian approach to classify individuals into groups using genotype data. Such data, also called multilocus genotype data, consists of several genetic markers whose variations are measured at a series of loci for each sampled individual. Their method is based on a parametric model (model-based clustering) in which there are K groups (where K may be unknown), each of which is characterized by a set of allele frequencies at each locus. Group allele frequencies are unknown and modeled by a Dirichlet distribution at each locus within each group. A MCMC algorithm is then used to estimate simultaneously assignment probabilities and allele frequencies for all groups. In such a model, individuals are assumed to be independent, which does not take into account their possible spatial proximity.

The main goal of this work is to introduce spatial prior models and to assess their role in accounting for the relationships between individuals. In this perspective, we propose to investigate particular Markov models on graphs and to evaluate the quality of mean field approximations for the estimation of their parameters.

Maximum likelihood estimation of such models in a spatial context is typically intractable but mean field like approximations within an EM algorithm framework, in the spirit of [9] will be considered to deal with this problem. This should result in a procedure alternative to MCMC approaches.

This first approach is based on traditionnal hidden Markov models for which a standard conditional independence assumption holds. Dependencies between individuals are described through spatial correlations between groups meaning that spatially close individuals are likely to belong to the same group. In a recent on going work joint with INRA Avignon, we tried to weaken the standard but somewhat irrealistic conditional independence assumption to describe dependencies at the observations level through a spatial correlation model inspired

[18] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

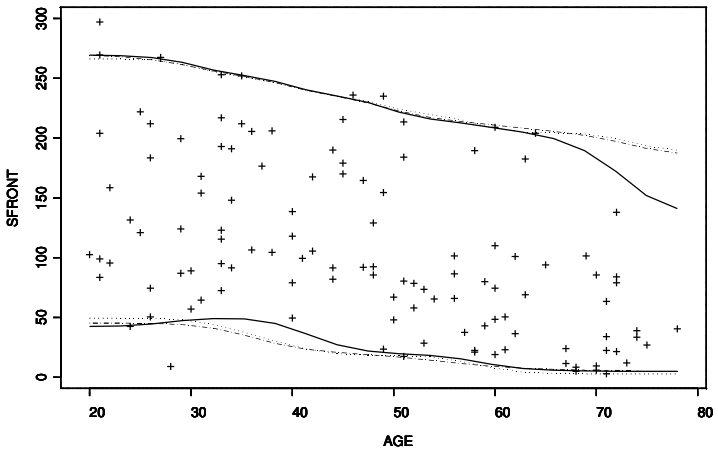


Figure 10: Reference curves for a women population: the sebum rate measured on the forhead of each woman is reported against the age. The curves show various estimates of the 10% reference curves bounding 90% of the population.

from geostatistical models [8].

Another goal of this work would then be to investigate such extension in the context of population genetics. In particular, mean field like approximations and more generally structural variational approximation techniques could again be considered to solve the consequent estimation problem.

6.2.3 Biophysical properties of women skin

Participant: Stéphane Girard.

The computation of reference curves (mentionned in Section 5.3.1) can be interpreted as a regression quantiles estimation problem, the explicative variable being the age of the subjects, for instance. When the covariate is unidimensional, we have developped a non-parametric kernel method to estimate these regression quantiles. The background of this research is the study of biophysical properties of women skin (see Figure 10) in collaboration with Chanel Research Center (CERIES). When the covariate is multidimensional, this approach has been combined with a semi-parametric dimension reduction method [28], the so-called SIR (Sliced Inverse Regression) method.

6.3 Reliability

6.3.1 An Aging model

Participant: Henri Bertholon.

Joint work with Gilles Celeux and Nicolas Bousquet from team SELECT, INRIA Futur.

In the reliability context, we are interested in lifetime data analysis. We have especially examined a simple competing risk model that may be viewed as a possible alternative to the standard Weibull model. In particular our model enables to take into account both accidental causes of failure and aging. The estimation of parameters is made by Maximum Likelihood and Bayesian inference. Moreover in order to discriminate between our model and Weibull (or exponential) models, a test procedure has been proposed. Finally different applications have been presented. The first one is related to the analysis of a prospective mortality table and the second one concerns the study of failure times for a fleet of 50 vehicles. In both cases, it is proved that our model is more adapted than the Weibull one.

7 Softwares

Our core techniques are usually made available under the forms of software libraries and binary code. The importance and usefulness of these libraries and binary code is twofold. First, we want to provide software in addition to our scientific developments and findings in order to avoid implementation errors by third parties. Second, the dissemination of both scientific results and their associated software packages will increase the impact of our work and will maximize its dissemination towards other academics and towards industry.

7.1 The EXTREMES freeware

Participant: Stéphane Girard.

This is 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 [29]. 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 available for Windows and, since August 2005, also for Linux. The current version 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 [30]. More generally, the statistics community has shown its interest in this software. It has been the subject of several communications and publications, among them:

- Journées de la SFDS, Lyon, France, 2003 [31];
- Extreme Value Analysis Conference, Aveiro, Portugal, 2004 [32];
- Extreme Value Analysis Conference, Gothenburg, Sweden, 2005 [33];
- La revue de Modulad, n 30, pp 53–60, 2003 [29];

- and various other publications related to its theoretical content.

Besides, several users had asked for the integration of new statistical methods related to bias reduction, modelling of the dependence between variables, censoring, etc. In addition, it is necessary to offer a wider field of application to this software. To this end, we are looking for financial support to develop a JAVA interface to obtain a software independent from the user platform. Also, we are considering the development of a web site including new sections such as FAQ, bug report, etc.

The targetted users include statisticians but also engineers. Other softwares on extremal events exist but most of them are developped for experts. None of them propose the same variety of features with the same accessibility.

7.2 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, written in C, is publicly available at: `http://mistis.inrialpes.fr/software/SEMMS.html`. A new version written in C++ is available upon request since November 2005.

Publications and communications related to the theoretical basis and the various methods made available in the package include:

- Pattern Recognition 2003 [9];
- IEEE Pami, 2003 [12];
- British Machine Vision Conference, Oxford, UK, 2005 [20];
- Applied Stochastic Models and Data Analysis Conference, Brest, France, 2005 [34].

The new version of the software will be publicly available soon with an appropriate user interface. In addition, new models and methods will be included related to new developments in Juliette Blanchet' PhD thesis.

8 Provisional programme of work and expected results

Our main mid-term goal (2 to 4 years) is to develop a general statistical formalism and a set of tools to analyse complex spatial data. We will focus mainly on two source of complexity by dealing with structured data or spatially organized data but in a non predictable regular way, and also with high dimensional data and various mixture of the two. We will provide tools to deal with irregularly located datapoints associated to high-dimensional features. We will give at least four illustrations of our techniques with one application regarding computer vision, object recognition and classification, one application in genomics, one application in population genetics and one application on remote sensing data.

Object recognition: particularity lies in the complexity/variability of the object classes, the high-dimension of the descriptors and the choice of their organizational model. This will be based on the Movistar project (see Section 10.1) and our collaboration with team Lear.

Genomics: particularity lies in the integration of the various type of available information and the question of how to weight these various sources. The data may also be high-dimensional but the treatement will be similar as above. Additional issues will appear due to the problem of interpreting the results. Specific problems also arise due to the difficulty in setting biological information/issues into statistically well formulated problems. The context is that of the IBN project (see Section 10.1) and our collaboration with team Helix.

Population genetics: particularity lies in the discrete nature of the data (eg. allele frequencies, Section 6.2.2) and the comparison/combination with Bayesian modelling. The choice of an organizational model also arises. The context is our collaboration with team TimB from TIMC laboratory.

Teledetection: a first data visualization stage/goal will be reached using the dimension reduction and segmentation techniques developped above. But space-time data requires specific treatement and additional techniques regarding curves analyses will be used to solve a second stage physical model inversion issue. The context is our collaboration with Laboratoire de Planétologie de Grenoble.

We will also focus on the issue of modelling and taking into account *high-level* a priori knowledge. Our approach will be based on integrated Markov models. We will go on investigating the use of such models for MRI analysis (see Section 6.1.6 and our collaboration with the INSERM/TIMC unit) and investigate their extensions to the modelling of perception (visual and auditory) in the context of a new European project POP coordinated by Radu Horaud from team MOVI/Perception (see Section 10.1).

In parallel, we will provide some of our techniques with theoretical results mainly about convergence and speed.

Also new softwares will be available either in the form of new improved version of already available ones or in the form of additional codes making new techniques available. New methods and new interface will be added in Extremes. A new package with new functionalities will complement the current SEMMS package. Also a new Matlab toolbox will be available for the visualization and analysis of time series of satellite images.

9 Positioning

9.1 Related INRIA teams

The activities of MISTIS regarding statistical methodology overlap at least four of the seven INRIA challenges through key-words such as *complex systems*, *information*, *applications to medicine and biology*.

MISTIS is classified in the INRIA thema *Cognitive Systems (A): Statistical modelling and machine learning*. The only other team focused on statistical methodology in this group is the team of G. Celeux, SELECT: *Model selection in statistical learning*. We have similar objectives but we differentiate from this team in that model selection is included in part of the issues we want to address but we do not aim at providing a general methodology to deal with this problem. We may actually be interested in using new methods proposed in SELECT. Besides, Markov and graphical models, the associated approximation techniques and the semi and non-parametric aspects we intend to develop are not included in SELECT objectives.

In the same group, team TAO focus on Machine Learning and Optimization. More contacts with this team involved in the learning community would be interesting for MISTIS.

In group *Cognitive Systems (B): Perception, indexing and communication for images and video*, we share interests with ARIANA and VISTA as regards the use of Markov models and wavelets in image analysis. However, although we currently deal with images in a number of applications, our aim is not to restrict to such data but to focus more on the idea of structure underlying various systems.

In this group, our main interaction is with teams LEAR and MOVI/Perception. Our statistical expertise is greatly enhanced through combination with their expertise in computer vision. We work closely with team LEAR through two co-advised PhD students and a common national project *ACI masses de données* [20, 2, 3, 35] (see Movistar in Section 10.1).

We interact with MOVI/Perception through pointwise collaborations [36] and a European project POP (see Section 10.1) which also includes the co-advising of a PhD student.

Other interactions would be interesting with team PRIMA, in group *Cognitive Systems (C)*, as regards mixture models for instance.

In group *Biological Systems (A)*, we interact with team HELIX through our participation to a Cooperative Research Initative project (ARC IBN, see Section 10.1) and through Matthieu Vignes' PhD thesis.

Two other teams with statistics background are ASPI and SISTHEM in the *Numerical Systems (C)* group. The first one deals like us with hidden Markov Models with special emphasis on hidden Markov chains (1-dimensional) while we focus more on 2-dimensional models. Also we use rather different techniques. Their approach is based on filtering and particle systems with focus on simulations and Monte Carlo methods while we are more interested in deterministic approximations. The second team focus on Structural Health Monitoring with statistical inference tools including identification, change detection, rejection methods rather different than ours.

In the same group, team IDOPT/MOISE is using statistical methods but not directly in

our domain of expertise.

In group *Numerical Systems (A)*, we have common interests with e-Motion regarding Bayesian networks. This team focus on Bayesian programming applied to robotics while we have a more formal approach of Bayesian techniques. As a matter of fact, we mentionned the idea of organizing common working groups with this team.

Eventually, in group *Numerical Systems (B)*, we share common probabilistic interests with team MESCAL. We may develop interactions with this team as regards structured models and simulation techniques. Both our teams are member of a local workshop (Programme Pluri Formations) including other academic members around common probabilistic tools.

9.2 National positioning

In the local context, we are closely related to the other statistics teams in Grenoble, namely teams SMS and LABSAD, through collaborations but also co-organized seminars and workshops. In addition MISTIS asked to be part of the new LJK laboratory (about 44 permanent researchers in the statistics department) that will gather locally various applied mathematics groups including the two previously mentionned and many others.

Possible interactions with these teams include functional inference, multi resolution analysis, dimension reduction techniques, Markovian and graphical models and reliability.

MISTIS' originality is mainly in our specific expertise in Markovian graphical models for spatial data, mixture modelling combined with interaction modelling and high-dimensional data analysis.

Outside the local context, other groups in France focus on research themes related to ours. These groups include an INRA team in the Biometry unit in Avignon (about 6 permanent researchers). It is specialized in spatial statistics and uses approches including Markovian models, Gibbs fields, stochastic simulation techniques, Monte Carlo Markov chains. Some of its research interests are closely related to ours, in particular as regards spatial data classification. We know some of its members very well and have common work and publication with two of its permanent members (see Section 5.2.3).

In Compiègne, the ASTRID team (14 permanent researchers) in the HEUDIASYC unit focus on pattern recognition and data analysis, image and signal processing using statistical approaches for discrimination and classification. Its expertise is somewhat in between ours and that of team Lear as regards learning techniques for computer vision. We all participate in the same ACI project (see Section 10.1).

The TSI (Signal and Image Processing) department at ENST in Paris and more specifically team TSAC including about 15 permanent researchers, has similar research topics such as those regarding hidden Markov chains and hidden Markov fields. Gersende Fort is now working in this department since June 2005.

We also have interactions with the SSB group gathering about 30 permanent researchers from INRA, Génopole Evry, INA-PG, universities of Rouen and Toulouse. Its focus is on statistics applied to genomic data and we have common interests in particular around the use of graphical models for such data. This group is also involved in ARC IBN (see Section 10.1).

Among university teams, it worthes mentionning the probability and statistics team from university Montpellier 2 (18 permanent researchers). Common interests include mainly functionnal estimatio. Stéphane Girard is a former member of this group and is still collaborating (recent co-publications) with two of its members.

In addition, large statistics groups in France are located in Toulouse and Paris. Of interest for us is the MAFIA group (12 permanent researchers) from university Toulouse 3 which focus on non parametric statistics, stochastic algorithms and dependencies on a theoretical point of view. We do not have for now collaboration with this group but it would certainly be interesting for MISTIS to initiate some. In Paris, the LSTA team (20 permanent researchers) from university Paris 6, include researchers working on non parametric statistics and bayesian statistics also on a more formal point of view. We have collaboration and co-publications with one of its member.

9.3 International positioning

In Europe, the statistics department (about 15 researchers) at Glasgow university in Scotland focus on topics similar to ours. Image and spatial data analysis is a significant part of their research, in particular through projects regarding image reconstruction, Markov models estimation, Missing data and variational approximations. Some of this departement members would be very relevant partners for MISTIS. We shared some work and a co-signed publication [13] with Professor Mike Titterington, head of the department.

Let us also mention the Center for Statistics and Applications (CEAUL) at Lisbon university in Portugal (about 30 permanent researchers). Common topics of interest include spatial statistics and geostatistics, classification and data mining.

In the United States, we are especially in touch with the statistics department of the university of Washington in Seattle (about 30 permanent researchers). Various common publications and visits to this department account for our fructful relationship with some of its members. There exist a number of significant statistics department in the States with which we have common research interest and whose research topics intersect with ours. Let us cite as an example Purdue and Berkeley (about 40 permanent researchers) universities.

In Australia, the Statistics program of the Mathematical Sciences Institut (MSI) at Australia National University (ANU) in Camberra gathers about 10 researchers. Its research interests overlap with ours as regard mixture models, spatial statistics, boundary estimation, discriminant analysis and dimension reduction.

10 Scientific collaborations

10.1 Contracts and grants

MISTIS got a Ministry grant (Action Concertée Incitative Masses de données) for a three-year project (2003-2006) 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. The PhD work of Juliette Blanchet and Charles Bouveyron are related to this project.

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. The PhD work of Matthieu Vignes is related to this project.

MISTIS is then involved in a European STREP proposal, named POP (Perception On Purpose) coordinated by Radu Horaud from INRIA team MOVI/Perception. The other partners are the universities of Osnabruck, Hospital Hamburg-Eppendorf, Coimbra and Sheffield. 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. A PhD student will be hired in June 2006 to work on this project.

10.2 Collaborations

We have been collaborating with some of the members of team ASTRID from unit Heudiasyc at UTC in Compiègne. In particular Christophe Ambroise and since July 2003 Yves Grandvalet with the beginning of the Movistar project.

Besides the Movistar partners, our research regarding dimension reduction in image analysis, is also joint work with Serge Iovleff from university of Lille I.

We collaborate with Denis Allard and Nathalie Peyrard from INRA, Avignon, on statistical methods and models to cluster geostatistical data (see Section 5.2.3).

We have joint work with Olivier Francois from team TimB in TIMC laboratory regarding spatial models for population genetics (see Section 6.2.2). The post-doc work of Chibiao Chen is part of this collaboration.

We are also involved in the PhD thesis of Benoit Scherrer co-advised by Catherine Garbay and Michel Dojat from INSERM and TIMC. Benoit's work regards MRI segmentation and analysis (see Section 6.1.6).

We are collaborating with Sylvain Douté and Etienne Deforas from Laboratoire de Planétologie de Grenoble on the analysis of remote sensing data. The post-doc work of Monica Benito is related to this collaboration.

We have joint work with Prof. Alexandre Nazin from Institute of Control Science in Moscow, Russia. Prof. A. Nazin has been visiting us regularly for periods of 2-3 months. The most recent common works [16, 37] deals with frontier estimation by linear programming (see section 5.3.1).

We have joint work with X and Y from the probability and statistics team of university Montpellier 2.

We also work with Armelle Guillou, from LSTA, university of Paris 6.

We have collaboration with Instituto de Sistemas e Robotica of Instituto Superior Tecnico in Lisbon, Portugal. Paulo Gonçalves spent two years (September 2003-September 2005) in this department and just started to co-advised Hugo Carrão, a PhD student from IGP, Lisbon.

We also have strong collaboration [26, 23, 24] with the Statistics department of university of Washington in Seattle. Florence Forbes visited the department on a regular basis (five visits) for periods of 3-4 months and was involved in the Model-based clustering and Applications workshop organized by Prof. Adrian Raftery.

10.3 Industrial contracts

The team does not currently have contracts with industrial partners. However, we give in this section some evidence of our efforts to create and develop such collaborations.

Our last contract was under the advising of C. Lavergne and was initiated in the former IS2 team, with the LCFR (Laboratoire de Conduite et Fiabilité des Réacteurs) of CEA/Cadarache/DER. It ended in 2005 and funded during three years the PhD thesis of J. Jacques on sensitivity analysis. Our main contact at CEA/Cadarache is N. Devictor and we are hopping to pursue this collaboration on different topics through a Master thesis and a co-advised PhD thesis on reliability and uncertainties statistical analysis. Another possibility we plan to investigate, with A. de Crecy from CEA/Grenoble, is the use of EM-like algorithms in the modeling of physical processes with unobserved states.

Our more recent efforts include a first meeting with Trixell in the Thales group (J.M. Vignolle) on statistical image analysis and a planned INRIA meeting with France-Telecom around massive data treatment.

References

[1] G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri. A component-wise EM algorithm for mixtures. *Journal of Computational and Graphical Statistics*, 10:699–712, 2001.

[2] C. Bouveyron, S. Girard, and C. Schmid. Analyse discriminante de haute dimension. Technical Report RR-5470, INRIA, 2005.

[3] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering in class-specific subspaces: Application to object recognition. Technical Report RR-XXXX, INRIA, 2005.

[4] C. Bouveyron, S. Girard, and C. Schmid. *Class specific subspace Discriminant analysis for high dimensional data*, volume to appear of *Lecture Notes in Computer Science*. Springer, 2006.

[5] C. Bouveyron, S. Girard, and C. Schmid. High Dimensional Discriminant analysis. *Communications in Statistics*, to appear, 2006.

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

[7] B. Schrerrer, M. Dojat, F. Forbes, and C. Garbay. Segmentation markovienne distribuée et coopérative des tissus et des structures présents dans des IRM cérébrales. In *RFIA*, Tours, France, 2006.

[8] D. Allard, F. Forbes, and N. Peyrard. Comparing two models for clustering geostatistical data. In *Working group on model-based clustering, Tenth anniversary Summer Session*, Univ. of Washington, Seattle, USA, July 2004.

[9] G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003.

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

[11] G. Fort and E. Moulines. Convergence of the Monte-Carlo EM for curved exponential families. *Annals of Statistics*, 31(4):1220–1259, 2003.

[12] F. Forbes and N. Peyrard. Hidden Markov model selection based on mean field like approximations. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 25(8), 2003.

[13] G. Celeux, F. Forbes, C.P. Robert, and M. Titterington. Deviance information criteria for missing data models. with discussion. *To appear in Bayesian Analysis*, 2005.

[14] S. Girard and L. Menneteau. Central limit theorems for smoothed extreme value estimates of point processes boundaries. *Journal of Statistical Planning and Inference*, 135(2):433–460, 2005.

[15] L. Gardes. *Estimation d'une fonction quantile extrême.* PhD thesis, Université Montpellier 2, octobre 2003.

[16] G. Bouchard, S. Girard, A. Iouditski, and A. Nazin. Nonparametric frontier estimation by linear programming. *Automation and Remote Control*, 65(1):58–64, 2004.

[17] S. Girard and S. Iovleff. Auto-associative models and generalized principal component analysis. *Journal of Multivariate Analysis*, 93(1):21–39, 2005.

[18] B. Chalmond and S. Girard. Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(5):422–432, 1999.

[19] J. B. Durand, P. Gonçalves, and Y. Guédon. Statistical inference for hidden Markov tree models and application to wavelet trees. *IEEE, Trans. on Signal Processing*, 52(9):2551–2560, 2004.

[20] J. Blanchet, F. Forbes, and C. Schmid. Markov random fields for textures recognition with local invariant regions and their geometric relationships. In *British Machine Vision Conference*, Oxford, UK, September 2005.

[21] P. Gonçalves, H. Carrão, A. Pinheiro, and M. Caetano. Land cover classification with Support Vector Machine applied to MODIS imagery. In *Proc. 25th EARSeL Symposium*, Porto, Portugal, June 2005.

[22] P. Oliveira, P. Gonçalves, and M. Caetano. Land cover time profiles from linear mixture models applied to MODIS images. In *Proc. 31st International Symposium on Remote Sensing of Environment*, Saint Petersburg, Russian Federation, June 2005.

[23] F. Forbes, N. Peyrard, C. Fraley, D. Georgian-Smith, D. Goldhaber, and A. E. Raftery. Region of interest selection and dynamic breast MRI data analysis using multivariate statistical methods for clustering and spatial segmentation. Technical Report RR-4249, Inria Rhône-Alpes, 2001.

[24] F. Forbes, N. Peyrard, C. Fraley, D. Georgian-Smith, D. Goldhaber, and A. E. Raftery. Model-based region of interest selection in dynamic breast MRI. Technical Report 472, Stat. dept, Univ. of Washington, 2004.

[25] F. Forbes, N. Peyrard, C. Fraley, D. Georgian-Smith, D. Goldhaber, and A. E. Raftery. Model-based region of interest selection in dynamic breast MRI. *Journal of Computer Assisted Tomography*, 2006. To appear.

[26] F. Forbes and A. E. Raftery. Bayesian morphology: Fast unsupervised bayesian image analysis. *Journal of the American Statistical Association*, 94(446):555–568, June 1999.

[27] F. Forbes and M. Vignes. Champs de markov cachés et fusion de données individuelles et pairées pour l'identification de groupes de gènes. In *JOBIM*, Lyon, France, Juillet 2005.

[28] A. Gannoun, S. Girard, C. Guinot, and J. Saracco. Sliced inverse regression in reference curves estimation. *Computational Statistics and Data Analysis*, 46(1):103–122, 2004.

[29] J. Diebolt, J. Ecarnot, M. Garrido, S. Girard, and D. Lagrange. Le logiciel Extremes, un outil pour l'étude des queues de distribution. *La revue de Modulad*, 30:53–60, 2003.

[30] J. Diebolt, M. Garrido, and S. Girard. A goodness-of-fit test for the distribution tail. In M. Ahsanullah and S. Kirmani, editors, *Topics in extreme values*. Nova Science, New-York, 2005. *A paraître*.

[31] J. Diebolt, J. Ecarnot, M. Garrido, S. Girard, and D. Lagrange. Le logiciel Extremes. In *35mes Journes de Statistique organises par la Socit Franaise de Statistique*, Lyon, France, 2003.

[32] M. Garrido, S. Girard, and J. Ecarnot. The Extremes software. In *Third International Symposium on Extreme Value Analysis*, page 27, Aveiro, Portugal, juillet 2004.

[33] J. Diebolt, M. Garrido, S. Girard, and 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.

[34] J. Blanchet, F. Forbes, and C. Schmid. Markov random fields for recognizing textures modeled by feature vectors. In *International Conference on Applied Stochastic Models and Data Analysis*, Brest, France, May 2005.

[35] C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. In *International Conference on Applied Stochastic Models and Data Analysis*, pages 526–534, Brest, France, May 2005.

[36] G. Dewaele, F. Devernay, R. Horaud, and F. Forbes. The alignment between 3D-data and articulated shapes with bending surfaces. 2006. European Conf. Computer Vision.

[37] G. Bouchard, S. Girard, A. Iouditski, and A. Nazin. Some linear programming methods for frontier estimation. *Applied Stochastic Models in Business and Industry*, 21(2):175–185, 2005.

[38] P. Gonçalves, C. Lenoir, C. Heymes, B. Swynghedauw, and C. Lavergne. Statistical modelling of cardiovascular data. an introduction to linear mixed models. Technical Report RR-5787, INRIA, 2005.

[39] N. Peyrard, F. Forbes, and 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.

[40] C. Bouveyron, S. Girard, and 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.

[41] C. Bouveyron, S. Girard, and 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.

[42] C. Bouveyron, S. Girard, and 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, September 2005.

[43] J. Blanchet, F. Forbes, and C. Schmid. Modèles markoviens pour l'organisation spatiale de descripteurs d'images. In *7e Conférence francophone sur l'Apprentissage Automatique, Presses Universitaires de Grenoble*, pages 113–126, Nice, France, Juin 2005.

[44] J. Blanchet, F. Forbes, and 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, Juin 2005.

[45] G. Celeux, F. Forbes, C. P. Robert, and M. Titterington. Deviance information criteria for missing data models. Technical Report RR-4859, Inria Rhône-Alpes, 2003.

[46] H. Bertholon. *Une modélisation du vieillissement.* PhD thesis, Université Joseph Fourier, Grenoble 1, 2001.

[47] G. Fort. *Contrôle explicite d'ergodicité de chaine de Markov: Applications à l'analyse de convergence de l'algorithme Monte-Carlo EM.* PhD thesis, Paris 6, 2001.

[48] S. Girard and P. Jacob. Extreme values and kernel estimates of point processes boundaries. *ESAIM: Probability and Statistics*, 2004. *A paraître.*

[49] L. Gardes and S. Girard. Asymptotic properties of a Pickands type estimator of the extreme value index. In F. Colombus, editor, *Focus on probability theory.* Nova Science, New-York, 2004. *A paraître.*

[50] S. Girard. A Hill type estimate of the Weibull tail-coefficient. *Communication in Statistics - Theory and Methods*, 33(2):205–234, 2004.

[51] S. Girard and P. Jacob. Extreme values and Haar series estimates of point process boundaries. *Scandinavian Journal of Statistics*, 30(2):369–384, 2003.

[52] S. Girard and P. Jacob. Projection estimates of point processes boundaries. *Journal of Statistical Planning and Inference*, 116(1):1–15, 2003.

[53] S. Girard. A nonlinear PCA based on manifold approximation. *Computational Statistics*, 15(2):145–167, 2000.

[54] A. Gannoun & S. Girard & C. Guinot & J. Saracco. Reference ranges based on non-parametric quantile regression. *Statistics in Medicine*, 21(20):3119–3135, 2002.

[55] A. Gannoun & S. Girard & C. Guinot & J. Saracco. Trois méthodes non paramétriques pour l'estimation de courbes de référence - application à l'analyse de propriétés biophysiques de la peau. *Revue de Statistique Appliquée*, L(1):65–89, 2002.

[56] M. Garrido. *Modélisation des événements rares et estimation des quantiles extrêmes, méthodes de sélection de modèles pour les queues de distribution.* PhD thesis, Université Grenoble 1, juin 2002.

[57] L. Gardes and S. Girard. Estimating extreme quantiles of Weibull tail-distributions. Technical Report RR-1065, LMC, 2004. `http://www.inrialpes.fr/is2/people/girard/RR1065.ps`.