INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team select

Model Selection and Statistical Learning

Futurs

THEME COG

Table of contents

1. Team

Team Leader
Pascal Massart [Professor Université Paris-Sud, HdR]

Team Vice-Leader
Gilles Celeux [DR2 INRIA, HdR]

Administrative assistant
Marie-Carol Lopes [TR partially until September 2006]
Gina Grisvard [TR partially since September 2006]

Staff member Inria
Jean-Michel Marin [CR2 INRIA detached from Université Paris 9, CR1 INRIA since September 2006]

Staff member Université Paris-Sud
Christine Kéribin [Assistant Professor]
Marie-Anne Poursat [Assistant Professor]

Staff member Université Paris 5
Marc Lavielle [Professor Université Paris 5, HdR]
Jean-Michel Poggi [Professor Université Paris 5, HdR]

Ph. D. student
Sylvain Arlot [MESR grant]
Jean-Patrick Baudry [MESR grant]
Nicolas Bousquet [INRIA grant]
Sohie Donnet [MESR grant]
Merlin Keller [CIFRE grant]
Marc Lavarde [CIFRE grant]
Cathy Maugis [MESR grant]
Bertrand Michel [CIFRE grant]
Marie Sauvé [MESR grant]
Vincent Vandewalle [MESR grant]
Nicolas Verzelen [MESR grant]

Post-doctoral fellow
Agnès Grimaud [Post Doc. fellowship since September 2006]

Associate Engineers
Anwulin Echenim [since September 2006]
Franck Nasse [since September 2006]

2. Overall Objectives

2.1. Model selection in Statistics

Our research domain is statistics. In the last decades, statistical methodology has received a lot of contributions. Many different methods and algorithms are available in current softwares of statistical learning. The user of these methods is facing the problem of choosing a relevant method for its data set and objective. The model selection problem is an important but difficult problem from both theoretical and practical point of views. Classical criteria of models selection, based on often unrealistic assumptions, are penalized minimum contrast criteria with fixed penalties. SELECT is aiming to provide efficient model selection criteria with data driven penalty terms. In this context, SELECT is expecting to improve the toolkit of statistical model selection criteria from both theoretical and practical aspects. Currently, SELECT is focusing its effort on variable selection in statistical learning, non linear regression models with random effects, hidden structure models and supervised classification. Its domains of application concern reliability, curves classification, phylogeny

analysis and classification in genetics. New developments of SELECT activities are concerned with applications in biostatistics (statistical analysis of fMRI data, population pharmacology) and population genetics.

3. Scientific Foundations

3.1. General presentation

We learned from the applications we treated that some assumptions which are currently used in asymptotic theory for model selection are often irrelevant in practice. For instance, it is not realistic to assume that the target belongs to the family of models in competition. Moreover, in many situations, it is useful to make the size of the model depends on the sample size which make the asymptotic analysis breakdown. An important aim of SELECT is to propose model selection criteria which takes these practical constraints into account.

3.2. A non asymptotic view for model selection

An important purpose of SELECT is to build and analyze penalized log-likelihood model selection criteria efficient when the number of models in competition grows to infinity with the number of observations. Concentration inequalities are a key tool for that purpose and lead to propose data-driven penalty choices strategies. A major issue of SELECT consists of deepening the analysis of data-driven penalties both from the theoretical and the practical side. There is no universal way of calibrating penalties but there are several different rather general ideas that we want to develop, including heuristics derived from the Gaussian theory, special strategy for variable selection and making use of resampling methods [9], [2], [31].

3.3. Taking into account the modelling purpose in model selection

Choosing a model is not only a difficult problem from the theoretical point of view. Model selection criteria have been conceived to answer the difficulty that the data probability distribution P is unknown. But, beyond technical difficulties which can occur when choosing a model, it can be fruitful to take into account the purpose of the model user to get reliable and useful models for statistical description or decision tasks. As noticed earlier, most of standard model selection criteria are assuming that P is belonging to one of the considered models without considering the modelling purpose. This point of view would be useful not only from the practical point of view, but also it could help to avoid or overcome theoretical difficulties. Moreover, taking into account the modelling purpose would produce flexible model selection criteria with data-driven penalties [11]. This point of view can be expected to be useful in supervised Classification and hidden structure models. Finally, it is worth to mention that an alternative Bayesian approach for taking the modelling purpose into account can be expected to be useful in that setting [14].

3.4. Bayesian model selection

The Bayesian approach to statistical problems is fundamentally probabilistic. A joint probability distribution is used to describe the relationships between all the unknowns and the data. Inference is then based on the posterior distribution, the conditional probability distribution of the parameters given the observed data. Beyond the specification of the joint distribution, the Bayesian approach is automatic. Exploiting the internal consistency of the probability framework, the posterior distribution extracts the relevant information in the data and provides a complete and coherent summary of post data uncertainty. Using the posterior to solve specific inference and decision problems is then straightforward, at least in principle. The SELECT team is interested in applications of this Bayesian approach for model uncertainty problems where a large number of different models are under consideration. The joint distribution is obtained by introducing prior distributions on all the unknowns, here the parameters of each model and the models themselves, and then combining them with the distributions for the data. Conditioning on the data then induces a posterior distribution of model uncertainty that can be used for model selection and other inference and decision problems. This is the essential idea and it can be powerful. However, two major challenges confront its practical implementation: the specification of the prior distributions and the calculation of the posterior: [1] and[16].

3.5. Nonlinear mixed effect models

Mathematical modelling of the dynamic processes involved in biological processes constitutes an important application in biostatistics. Mixed effect models are very useful for modelling the variability within a population of these dynamic processes. Several statistical issues can be studied related to these models, such as parameter estimation, model selection (covariate model through the specification of fixed effect structure, covariance model for random effects), models defined by Ordinary or Stochastic Differential Equations, left censored models, as well as design optimization for the trial itself [23], [34], [34], [29] and [19].

4. Application Domains

4.1. General presentation

SELECT aims to produce methodological contributions in Statistics. For this very reason, the members of SELECT are involved in applications. We are considering that applications are important to provide us interesting practical problems for which there is the need of innovative methodologies. Most of the applications we are involved concern contracts with industrial partners (for instance our activities in reliability), and some of them concern more academic collaborations (as our activity in phylogeny).

4.2. Curves classification

An increasing interest is now evident in the field of classification and regression for complex data as curves, functions, spectra, time series. Such questions naturally arise when each observation consists of values of explanatory variables which are not scalar valued but of functional nature. Classical questions widely examined in Data Analysis are now revisited to take into account and advantage (if possible) of the functional nature of the data and to define original strategies [6]. Such questions are now related to a well identified domain called functional data analysis. Various applied problems strongly motivate this interest like longitudinal studies, analysis of fMRI data, spectral calibration, ...

We are focusing on classification problems with a particular emphasis on clustering (unsupervised classification) ones. In addition to classical questions like the choice of the number of clusters, the norm to measure the distance between two observations, or the vectors to represent clusters, a crucial problem naturally arises: due to the functional nature of the data, the computational effort needed is quickly huge and efficient algorithm as well as anytime algorithms are of interest.

4.3. Reliability

An important theme that SELECT considers is *aging modelling*. This research is done thanks to a contract with EDF-DER *Fiabilité des Composants et Structures* group. Most of the French nuclear park is approaching forty years which is the warranty age of good running. EDF is interested to examine the possible extension of use of nuclear material components beyond forty years and has planned studies to analyze durability of nuclear components and aging mastership. The collaboration of SELECT with EDF takes place in this framework.

The other theme of research in which SELECT is involved concerns changes in a reliability process. It comes from a contract with Altis firm. During the last five years, Altis has drastically changed its production process of chips. Indeed half of the production is nowadays made with brass connexions instead of aluminum connexions. This makes the usual reliability model irrelevant. Some abrupt change of the reliability behavior is suspected. We are working on the selection of a good model fitting data.

4.4. Phylogeny

Phylogeny is concerned with designing evolutionary trees between species from aligned nucleotide sequences . More precisely, a nucleotide sequence being an ordered set of sites taking value in a finite set E (for instance, $E = \{A, C, G, T\}$), the problem is to reconstruct the topology of the evolutionary tree between the species from aligned sequences for the considered species, and to estimate the tree parameters (branches length) as well as the parameters of the evolutionary model. Our research in this domain is twofold. First we are working on a model selection approach from a semi parametric graphical model whose parameters to be estimated are the topology, branches lengths and mutation rate of the evolutionary tree. Secondly, we are working on the *covarion* model. For this model, a site can change of behavior along the evolutionary tree according to two hidden states, active (ON) or nonactive (OFF). In this research, we are interested to compare non nested models.

4.5. Population genetics

SELECT develops new methods of statistical inference on molecular data obtained from population samples. Some of these methods are aimed at treating complex evolutionary scenarii, including several populations related by phylogenetic trees, with possible admixture and/or migration. Other methods will explicitly take into account the spatial distribution of samples. Inference concerns the parameters of these scenarii, which mainly characterize the population demographic history and the mutation model of markers. The explicit use of geographic information allow a more efficient characterization of evolutionary episodes poorly analyzed by existing methods, such as bioinvasions or shifts of species distribution areas due to global climatic changes. The analysis of complex scenarii will combine two algorithms: an Importance Sampling algorithm to estimate the data likelihood under a given scenario and with given values of parameters and a second algorithm (to be determined) to explore efficiently the parameter space.

4.6. Neuroimaging

A collaboration of SELECT with the SHFJ (Service Hospitalier Frédéric Joliot, CEA) concerns the statistical analysis of fMRI (functional Magnetic Resonance Imaging) time series. The aim of this research is to determine which parts of the brain are activated by different types of stimuli. A model selection approach is useful to avoid "false-positive" detections.

4.7. Population pharmacology

Pharmacokinetic (PK) studies (studies investigating the dose-concentration relationships of drugs) show for many drugs a large variability of pharmacokinetic parameters between individuals. Pharmacokinetic parameters describe processes such as absorption, diffusion and metabolism of drugs. The so-called "population PK approach" has been developed to characterise and quantify this variability, and is also applied to the study of pharmacodynamics (studies investigating the concentration-effect relationships of drugs). We have developed a complete methodology for the analysis of PK/PD data using a maximum likelihood approach.

An important application is the study of anti-HIV treatment. The efficiency of antiretroviral treatments, whether in HIV or hepatitis B or C pathologies, is quantified by the decrease in viral loads. Models have been developed to describe the time-course of this decrease through a system of ODE, taking into account the physiology of viral replication and the action mechanisms of the different therapeutic options. There is a large inter-patient variability in these pathologies, and the joint study of viral load decrease through mixed effect models in a set of patients provides a better understanding of differences in the response to treatment.

5. Software

5.1. MIXMOD software

Keywords: *Mixture model*, *cluster analysis*, *discriminant analysis*.

Participants: Gilles Celeux [Correspondant], Anwulin Echenim.

MIXMOD is developed with Christophe Biernacki, Florent Langrognet (Université de Franche-Comté) and Gérard Govaert (Université de Technologie de Compiègne). MIXMOD (MIXture MODelling) software fits mixture models to a given data set with either a clustering or a discriminant analysis purpose. A large variety of algorithms to estimate the mixture parameters are proposed (EM, Classification EM, Stochastic EM) and it is possible to combine them to lead to different strategies in order to get a sensible maximum of the likelihood (or completed likelihood) function [8]. Moreover, different information criteria for choosing a parsimonious model (the number of mixture component, for instance), some of them favoring either a cluster analysis or a discriminant analysis view point, are included. Many Gaussian models for continuous variables and multinomial models for discrete variable are available. Written in C++, MIXMOD is interfaced with SCILAB and MATLAB. The software, the statistical documentation and also the user guide are available on the Internet at the following address http://www-math.univ-fcomte.fr/mixmod/index.php. This year the Version 2.0 of MIXMOD including the mutinomial mixture models for treating qualitative variables has been made available. This new version includes specific graphical tools to display the results of mixture analysis with qualitative data. An expert engineer Anwuli Echenim has been hired to continue to improve the performances of this software which is already one of the most complete and rapid sofware on mixture analysis.

5.2. MONOLIX software

Keywords: *Non linear models*, *mixed effects*.

Participants: Marc Lavielle [Correspondant], Franck Nassé.

The MONOLIX group (Modèles Non Linéaires à Effets Mixtes), is chaired by France Mentré (INSERM-P7) and Marc Lavielle. This multi-disciplinary group, born in October 2003, has been meeting every month to exchange and develop activities in the field of mixed effect models. The group is actively engaged in producing a software to implement the methodology proposed. The new version of the MONOLIX software (http://www.math.u-psud.fr/~lavielle/monolix) is supported by Johnson & Johnson Pharmaceutical Research & Development. Marc Lavielle has presented the software in several occasions:

- PAGE meeting, Bruges, June 2006,
- Johnson & Johnson Pharmaceutical Research & Development , New-Jersey, October 2006,
- EUROBIO, Paris, October 2006,
- Pfizer, Sandwich (UK), December 2006.

We have obtained from INRIA FUTURS an ODL (Opération Développment Logiciel) to hire a engeneer (Franck Nassé). The aim of this ODL is to develop a new cross-platform C++ version of the MONOLIX software.

6. New Results

6.1. Model selection in Regression and Classification

Participants: Sylvain Arlot, Jean-Patrick Baudry, Gilles Celeux, Jean-Michel Marin, Pascal Massart, Cathy Maugis, Bertrand Michel, Jean-Michel Poggi, Marie Sauvé, Christine Tuleau.

In collaboration with Marie-Laure Martin (INRA), Gilles Celeux and Cathy Maugis [50] developed a variable selection procedure for model-based clustering which can be regarded as an improvement of a method proposed by Raftery and Dean (2006). The variable selection problem is recast as a global model selection problem, solved by comparing approximate Bayes factors. The procedure selects simultaneously the number of clusters, the form of the Gaussian mixture, the relevant variables for the clustering and the subset of relevant variables explaining irrelevant variables by linear regression. Encouraging performances for clustering transcriptom data have been obtained .

Cathy Maugis and Bertrand Michel started a theoretical work for selecting relevant variables by Gaussian mixture models where the mean vectors have the same values for some components. They aim to select the best model using a penalized criterion. This work is motivated by two practical problems: clustering of transcriptome data [50] and curve classification applied on oil production [51].
Jean-Patrick Baudry and Gilles Celeux have started a research to investigate the theoretical properties of ICL criterion (Biernacki, Celeux and Govaert, 2000) heuristically well sound to select a mixture model focusing on a classification purpose.
Jean-Patrick Baudry, Gilles Celeux, Jean-Michel Marin, Pascal Massart, Cathy Maugis and Bertrand Michel started a work to investigate the behavior of the "slope heuristic": In many contexts, the graph of the log-likelihood against the model complexity becomes almost linear with the complexity. Penalizing the log-likelihood by twice the slope of this linear part is advocated in [2]. The interest of the data driven penalty is investigated from simulations. First experiments are quite encouraging and we have the idea to use this heuristics to penalize a classification likelihood rather than the likelihood.

Marie Sauvé [5], [52] has considered the problem of choosing an histogram estimator of a regression function. The non asymptotic approach of model selection via penalization developped by Birgé and Massart is adopted but the observations are not assumed to be gaussian variables. A collection of partitions of $\mathcal{X}$, with possibly exponential complexity, and the corresponding collection of histogram estimators is considered. A penalized least squares criterion which selects a partition whose associated estimator performs approximately as well as the best one is proposed.
Marie Sauvé and Christine Tuleau [67] studied variable selection through CART method, both in the regression and binary classification frameworks. They propose an automatic and exhaustive procedure which relies on the use of the CART algorithm and model selection via penalization. This theorical work aims at determining adequate penalties allowing to get oracle type inequalities. A simulation study completes the theoretical results.

Sylvian Arlot and Pascal Massart defined a new model selection procedure in regression with resampling penalization, based on Efron's bootstrap and penalization. It is stated in a general form, including some random hold-out method (which is a kind of cross-validation), and could be used in many frameworks. In the case example of regression on histograms, their procedure is proved to satisfy an oracle inequality with constant almost one with high probability. It is also adaptive to the Hölder regularity of the target function. Numerical experiments show that the procedure is competitive with classical methods such as Mallow's C_p or cross-validation.

In collaboration with Christian Robert (Université Paris Dauphine), Gilles Celeux and Jean-Michel Marin [16] have considered Bayesian variable selection in linear regression. All its aspects have been studied in order to provide a precise and efficient userguide. The informative and noninformative cases has been analysed. In the informative case, it is suggested to choose the Zellner G-prior on the full model and to derive compatible prior distributions for each sub-model. In the noninformative case, it is shown that, if a Zellner weakly informative prior is used, the model posterior probabilities are sensitive to the choice of an hyperparameter. Consequently a new Zellner hierarchical prior is proposed. The use of this prior is shown to outperform penalized likelihood criteria in an explicative point of view. Finally, computational aspects are considered when the number of variables is large and, it is shown that the Gibbs sampling do the job quite well. Rigth now, the authors compare their approach in terms of prediction ability with the adaptive Lasso.

In collaboration with Servane Gey (Université Paris V), Jean-Michel Poggi [22] considered the AdaBoost like algorithm for boosting CART regression trees. The boosting predictors sequence is analysed on various data sets and the behaviour of the algorithm is investigated. An instability index of a given estimation method with respect to some training sample is defined. Based on the bagging algorithm, this instability index is then extended to quantify the additional instability provided by the boosting process with respect to the bagging one. Moreover, the ability of boosting to track outliers and to concentrate on hard observations is used to explore a nonstandard regression context.

In collaboration with Nathalie Cheze (Université Paris-Sud), Jean-Michel Poggi [41] proposed a procedure for detecting outliers in regression problems based on information provided by boosting trees. The key idea is to select the most frequently resampled observation along the boosting iterations and reiterate boosting after removing it. The procedure is noise distribution free. A lot of well-known benchmark data sets are considered and a comparative study against two classical competitors highlights the value of the method.

6.2. Model selection for high-dimensional graphical models

Participants: Pascal Massart, Nicolas Verzelen.

Pascal Massart and Nicolas Verzelen [53] aim at defining adaptative estimation procedures for graphical models. These models appear in various fields such as spatial statistics, image analysis or microarray data. If the underlying graph is known, efficient estimators for the parameters exist. However, selecting the graph is a more difficult problem. Pascal Massart and Nicolas Verzelen introduced a penalized criterion based on Mallows heuristic in order to select the graph when the underlying process is stationnary.

6.3. Statistical learning methodology and theory

Participants: Gilles Celeux, Jean-Michel Marin, Pascal Massart, Laurent Zwald.

In collaboration with Christian Robert (Université Paris Dauphine) and Mike Titterington (University of Glasgow, Scotland), Gilles Celeux and Jean-Michel Marin [49] developed a probabilistic model of the k-nearest neighboor classifier. It has been shown that it is possible to design a Potts-like model answering the theoretical flaw of the recent model of Holmes and Adams (2002). The interest of such a probabilistic view of the k-nearest neighboor classifier is to allow to consider well ground variable selection procedures. A special attention has been paid on Bayesian inference. In particular, we are investigating the possibility of estimating the unknown normalizing constant of the model via *path sampling* and *perfect sampling* to avoid the pseudolikelihood approximation which could be poor as illustrated with numerical experiments.

In collaboration with Christophe Biernacki (Université de Lille) and Gérard Govaert (UTC Compiègne), Gilles Celeux continues Bayesian analysis of latent class models to analyse multivariate multinomial discrete data sets. In particular, a predictive approach for clustering approach has been developed. This fully non in informative approach is investigated from a statistical view point. to select a sensible number of mixture components by using the integrated likelihood of a model. But it leads also to algorithmic research since this criterion is difficult to optimize. And, a collaboration with Marc Schoenauer and Damien Tessier from TAO team (INRIA Futurs) leads to compare their evolutionary algorithms with Monte Carlo Markov Chains algorithms. The performance of evolutionary algorithms are promising, but extensive experiments are scheduled to compare the algorithms more precisely.
In collaboration with Christophe Biernacki (Université de Lille) and Gérard Govaert (UTC Compiègne), Gilles Celeux started a research on semi-supervised classification with the aim to get new and general routines in the software MIXMOD to deal with semi-supervised classification. This reserch area is the subject of the thesis of Vincent Vandewalle started in October 2006.

In collaboration with Gilles Blanchard (University of Berlin, Germany), Laurent Zwald [54] developed his works on the Kernel Projection Machine (KPM). They proposed a nonasymptotic statistical analysis of Kernel-PCA with a focus different from the one proposed in previous work on this topic. They derived an upper bound of the error rate depending on the spacing between eigenvalues but not on the dimensionality of the eigenspace.

6.4. Adaptive importance sampling schemes

Participant: Jean-Michel Marin.

In collaboration with Christian Robert (Université Paris Daupine) and Antonietta Mira (University of Varese, Italy), Jean-Michel Marin developed new adaptive importance algorithms [48]. These are particular Population Monte Carlo schemes especially adapted to multivariate targets. Some of these algorithms do not require any tuning parameter. In collaboration with Jean-Marie Cornuet (INRA), the power of some of these schemes has been tested on population genetics models.

In collaboration with Roberto Casarin (University of Brescia, Italy), Jean-Michel Marin compared three regularized particulars filters in an on-line data processing context, in terms of parameter estimation and filtering ability. The Bayesian paradigm and the stochastic volatility model are considered. It is shown that the Regularized Auxliary Particle Filter (R-APF) outperforms the Regularized Sequential Importance Sampling (R-SIS) and the Regularized Sequential Importance Resampling (R-SIR).

6.5. Reliability

Participants: Nicolas Bousquet, Gilles Celeux, Romain François, Marc Lavarde, Pascal Massart, Jean-Michel Marin.

In the framework of a contrat with EDF concerning reliability, Nicolas Bousquet, Gilles Celeux and Jean-Michel Marin have:

- Design prior families for eliciting expert opinion for Bayesian inference in reliability. The discussion between the statistician and the expert is the focusing point of the study. They obtain a simple way of eliciting prior knowledge, when Weibull distributions are used for modelling aging lifetimes: [36] and [37].
- Propose and study DAC (Data Agreement Criterion) criterion for checking the consistency of expert opinion with respect to data in the context of subjective Bayesian analysis, [38] and [39]. Efficient approximations of DAC criterion has been proposed for practical analysis. Moreover, it has been shown that DAC criterion can be used as a sensible tool for calibrating prior subjective distributions [58].
- Define a methodological Bayesian tool for analyzing nuclear component lifetimes [3].

Gilles Celeux and Romain François have adapted and experimented the SAEM algorithm of Marc Lavielle to inverse problems occuring with deterministic models with random errors. This preliminary work allowed to discover identifiability problems which could be solved by imposing constraints or Bayesian analysis [43].

Accelerated life test (ALT) are widely used by manufacturers. The goal of ALT is getting reliability information in a very short time and having a better knowledge on the failure mechanism. The principle of ALT is to run the life test in a more severe than usual environment. In the framework of a contract with Altis and in collaboration with Patrick Pamphile (Orsay), Marc Lavarde and Pascal Massart [45] have adapted and applied the penalized model selection criterion of Birgé-Massart for an accelerated lifetime test problem.

6.6. Classification in genetics

Participants: Gilles Celeux, Cathy Maugis.

In collaboration with researchers of URGV (Evry Genopole) and Marie-Laure Martin (INRA), Gilles Celeux and Cathy Maugis made use of Gaussian mixture models to extract groups of coexpressed *Arabidopsis thaliana* genes. These models allow to take into account the existence of missing data and some priori biological information. For instance, we impose a cluster with a null mean and a spherical variance matrix can be imposed to take into account the existence of many no differential expressed genes in each experiment. Moveover, to improve the clustering and make easier the biological interpretation, a variable selection procedure has been designed [50].

6.7. Curves classification, denoising and forecasting

Participants: Pascal Massart, Bertrand Michel, Jean-Michel Poggi, Christine Tuleau.

In collaboration with Mina Aminghafari (Université Paris-Sud) and Nathalie Cheze (Université Paris-Sud), Jean-Michel Poggi [6] proposed a multivariate extension of the well known wavelet denoising procedure widely examined for scalar valued signals. It combines a straightforward multivariate generalization of a classical one and principal component analysis. This new procedure exhibits promising behavior on classical benchmark signals and the associated estimator is found to be near minimax in the one-dimensional sense, for Besov balls.

In collaboration with Mina Aminghafari, Jean-Michel Poggi [55] considered wavelets in time series, focusing on statistical forecasting purposes. A method estimating directly the prediction equation by direct regression has been studied and extended. The new variants are used first for stationary data, possibly contaminated by a deterministic trend.

In collaboration with Magalie Fromont (Université de Rennes), Chritine Tuleau [21], [44] has studied the k-nearest neighboor method for functional data. For functional data, the procedure consists of applying standard kNN on the projections of the data in a suitable space of dimension d. The procedure involves to select the dimension d and the number of neighbors k.

Hubbert's classical method of modelling oil production is based on fitting curve production with a logistic or Gaussian curve. In reality, bell curves sometimes correctly fit global production, but until now no rigorous explanation of this phenomenon has been given. Is is reasonable to think that the shape of the basin profile can be explained by the production dynamics of its individual fields. Pascal Massart and Bertrand Michel [51] propose a probabilistic model of oil production in a homogeneous geological zone.

6.8. Bayesian estimation and model selection

Participant: Jean-Michel Marin.

In collaboration with Pierre Druilhet (ENSAI, Rennes), Jean-Michel Marin [61] proposed a new version of MAP estimators and HPD credible sets. In the special case of non-informative prior, the new MAP estimators coincide with the equivariant frequentist ML estimators. They also proposed several adaptations when nuisance parameters are present.

In collaboration with Guido Consonni (University of Pavia, Italy), Jean-Michel Marin [17] studied the mean-field variational Bayesian inference. The behavior of this approach in the setting of the Bayesian probit model is illustrated. It is shown that the mean-field variational method always underestimates the posterior variance and provides poor approximation for small sample sizes.

In the Bayesian paradigm, Jean-Michel Marin [62] considered compatible prior distributions for model selection in a Bayesian setting. The idea that two priors are most compatible when the corresponding marginal distributions of the observations are closest to each other is developed.

In collaboration with Selima BenMansour, Elyes Jouini, Clotilde Napp and Christian Robert (all from Université Paris Dauphine), Jean-Michel Marin [47], [57] estimated on real dataset the average level of pessimism weighted by the risk tolerance. Its estimation leads to a nontrivial statistical problem. It was assumed that individuals have true unobservable characteristics and that their answers are noisy realizations of these characteristics. The Bayesian paradigm has been adopted and an hybrid MCMC approximation method used.

6.9. Neuroimaging, Statistical analysis of fMRI data

Participants: Sophie Donnet, Merlin Keller, Marc Lavielle.

Marc Lavielle and Sophie Donnet have tested a flexible model that allows for the variation of the magnitude of the Hemodynamic Response Function (HRF) with time in a Bayesian framework. Under this model, the magnitude of the HRF evoked by a single event may vary across occurrences of the same type of event. This model is tested against a simpler model with a fixed magnitude using information theory. They developed an EM algorithm to identify the event magnitudes and the HRF. They tested this hypothesis on a series of 32 regions of interest and find that the more flexible model is better than the usual model in most cases [18].

A collaboration of SELECT with the SHFJ (Service Hospitalier Frederic Joliot, CEA) concerns the statistical analysis of fMRI time series. In general, a convolution model is used to described the fMRI data. However, such models suffer from a lack of biological basis. Recently, physiological models have been introduced to understand the links between the neuronal activity and the hemodynamic phenomena. The BOLD signal measured by the MRI scanner is then descibed as the nonanalytical solution of a differential system. The input of this model are the neuronal efficiencies. Sophie Donnet and Marc Lavielle proposed to test a model

allowing a new variability in the neuronal efficiencies. Under this model, the neuronal efficiencies may vary across the type of stimuli. This model is tested against a simpler model with a fixed neuronal efficiencies for all the stimuli using information theory [4], [42]. Moreover, in collaboration with Adeline Samson, Sophie Donnet developped a general method to estimate the parameters for regression models defined by differential system [19]. They tested this model on a real data set extracted from the primary visual cortex. They found that the more flexible model is better than the usual model.
Merlin Keller began its PhD in October 2006 under the supervision of Alexis Roche (CEA, SHFJ) and Marc Lavielle.

6.10. Nonlinear mixed efects model

Participant: Marc Lavielle.

As previously explained, the MONOLIX group, co-chaired by Marc Lavielle develops activities in the field of mixed effect models. This group involves scientists with varied backgrounds, interested both in the study and applications of these models. Several papers have been produced [23], [34], [65], [29].

7. Contracts and Grants with Industry

7.1. Contracts with EDF

Participants: Nicolas Bousquet, Gilles Celeux.

- SELECT has a contract with EDF regarding durability of nuclear components and aging mastership.
- SELECT has a contrat with EDF regarding modelling uncertainty in deterministic models.

7.2. Other contracts

Participants: Marc Lavarde, Pascal Massart, Bertrand Michel, Marie Sauvé.

- SELECT has a contract with Altis (CIFRE grant of Marc Lavarde) regarding accelerated lifetime tests in the production process of chips.
- SELECT has a contract with IFP (CIFRE grant of Bertrand Michel) on modelling exploitation process of a petrol basin. Purposes of this work are the classification of production profiles and developing model selection tools in the context of Poisson process.
- The thesis of Marie Sauvé [5] is supported by Rhodia.

8. Other Grants and Activities

8.1. National Actions

SELECT is animating a working group on model selection and statistical analysis of genomics data with the Biometrics group of Institut Agronomique Nationale Paris-Grignon (INAPG).

Pascal Massart and Jean-Michel Marin are organizing a working group at ENS (Ulm) on Statistical Learning. This year the group focused interest on large dimension problems and Graphical Models. Most of SELECT members are involved in this working group.

8.1.1. MONOLIX Group

Participants: Sophie Donnet, Marc Lavielle.

The MONOLIX group chaired by Marc Lavielle and France Mentré (INSERM) is a multidisciplinary group, that exchanges and develops activities in the field of mixed effect models. It involves scientists with varied backgrounds, interested both in the study and applications of these models: academic statisticians (theoretical developments), researchers from INSERM (applications in pharmacology) and INRA (applications in agronomy, animal genetics and microbiology), and scientists from the medical faculty of Lyon-Sud University (applications in oncology).

8.1.2. Action incitative DataHighDim

Participant: Gilles Celeux.

This ACI started in September 2003. Partners of ACI DataHighDim are laboratory CLIPS of UJF and laboratory LIS, INPG in Grenoble, SELECT team of INRIA, laboratory DICE, UCL in Louvain la Neuve and laboratory LDG, CEA Bruyères le Châtel. DataHighDim is concerned with exploratory and decisional analysis in high dimensions. This year with Eugène Ndong Guema (Université de Yaoundé) has continued the work of Guillaume Saint Pierre on supervised classification using distance tables.

8.2. European actions

Gilles Celeux and Pascal Massart are members of the PASCAL (Pattern Analysis, Statistical Learning and Computational Learning) network. Jean-Michel Marin spent two week in Pavia in the statistical department of Pavia University. He gave a conference during his stay.

9. Dissemination

9.1. Scientific Communauty animation

- Gilles Celeux is editor-in-chief of *Statistics and Computing*.
- Pascal Massart is associated editor of *Annales de l'IHP*, *Journal of the European Mathematical Society*, *Journal de la SFDS* and *ESAIM Proceedings*.

- Gilles Celeux was invited speaker at ECAIS 2006 (University Paris 5) in November 2006 and to the "Jean-Pierre Fénelon Cycle de Conférences" of INA.
- Pascal Massart was invited speaker at the 9th International Vilnius Conference on Probability Theory and Mathematical Statistics (Vilnius, Lithuania) in June 2006.
- Pascal Massart was invited speaker at the XXVI European Meeting of Statisticians in (Turon, Poland) in July 2006.

- Gilles Celeux has chaired the evaluation council of "Unité Jouy-en-Josas du département MIA de l'INRA".
- Marc Lavielle has organised "Journée Statistique et Santé : statisticiens, biostatisticiens et médecins se rencontrent" (Université Paris 5) in May 2006.
- Marc Lavielle is "Chargé de Mission, DSPT1 Mathématiques et leurs interactions, Mission Scientifique Technique et Pédagogique, Ministère délégué à la Recherche et aux Nouvelles Technologies".
- Jean-Michel Marin is the head of the council of the french statistical society.
- Pascal Massart is member of the scientific council of Euradom and of the working group on "le rôle des mathématiques dans le monde contemporain" of the french *Académie des Sciences*.
- Jean-Michel Poggi has been member of the Progam committee of Journées MAS Lille 2006.

9.2. Teaching

Pascal Massart is responsible of the M2 “Modélisation stochastique et statistique” of Orsay. All the SELECT members are teaching in various courses of different universities.

10. Bibliography

Year Publications

Books and Monographs

[1] J.-M. MARIN, C. ROBERT. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer texts in Statistics, Springer, New York, 2007.

[2] P. MASSART. *Concentration inequalities and model selection*, Lecture Notes in Mathematics, vol. 1896, Springer-Verlag, 2007.

Doctoral dissertations and Habilitation theses

[3] N. BOUSQUET. *Une méthodologie d'analyse bayésienne pour la prévision de la durée de vie de composants industriels*, Ph. D. Thesis, Université Paris-Sud, 2006.

[4] S. DONNET. *Inversion de données IRMF. Estimation et sélection de modèles*, Ph. D. Thesis, Université Paris-Sud, 2006.

[5] M. SAUVÉ. *Sélection de modèles en régression non gaussienne. Applications à la sélection de variables et aux tests de survie accélérés*, Ph. D. Thesis, Université Paris-Sud, 2006.

Articles in refereed journals and book chapters

[6] M. AMINGHAFARI, N. CHÈZE, J.-M. POGGI. *Multivariate denoising using wavelets and principal components*, in "Computational Statistics & Data Analysis", vol. 50, 2006, p. 2381–2398.

[7] H. BERTHOLON, N. BOUSQUET, G. CELEUX. *An alternative competing risk model to the Weibull distribution for modelling aging in lifetime data analysis*, in "Lifetime Data Analysis", to appear, 2006.

[8] C. BIERNACKI, G. CELEUX, G. GOVAERT, F. LANGROGNET. *Model–based cluster analysis and discriminant analysis with the MIXMOD software*, in "Computational Statistics & Data Analysis", vol. 51, 2006, p. 587–600.

[9] L. BIRGÉ, P. MASSART. *Minimal penalties for Gaussian model selection*, in "Probability Theory and Related Fields", to appear, 2006.

[10] G. BLANCHARD, P. MASSART. *Discussion on the 2004 IMS Medallion Lecture*, in "Annals of Statistics", vol. 34, n$^{\text{o}}$ 6, 2006.

[11] G. BOUCHARD, G. CELEUX. *Selection of generative models in classification*, in "IEEE Trans. on PAMI", vol. 28, 2006, p. 544–554.

[12] I. BRITO, G. CELEUX, A. FERREIRA. *Combining methods in supervised classification: a comparative study on discrete and continuous problems*, in "REVSTAT", vol. 4, 2006, p. 1–12.

[13] G. CELEUX, F. CORSET, A. LANNOY, B. RICARD. *Designing a Bayesian network for preventive maintenance from expert opinion in a rapid and reliable way*, in "Reliability Engineering & System Safety", vol. 91, 2006, p. 772–777.

[14] G. CELEUX, F. FORBES, C. ROBERT, M. TITTERINGTON. *Deviance information criteria for missing data models (with discussion)*, in "Bayesian Analysis", vol. 1, n^{o} 4, 2006, p. 651–705.

[15] G. CELEUX, J.-M. MARIN, C. ROBERT. *Iterated importance sampling in missing data problems*, in "Computational Statistics & Data Analysis", vol. 50, n^{o} 12, 2006, p. 3386–3404.

[16] G. CELEUX, J.-M. MARIN, C. ROBERT. *Sélection bayésienne de variables en régression linéaire*, in "Journal de la Société Française de Statistique", vol. 147, n^{o} 1, 2006, p. 59–79.

[17] G. CONSONNI, J.-M. MARIN. *Mean field variational Bayesian inference for latent variable models*, in "Computational Statistics & Data Analysis", to appear, 2006.

[18] S. DONNET, M. LAVIELLE, J.-B. POLINE. *Are fMRI event related response constant in time?*, in "NeuroImage", vol. 31, n^{o} 3, 2006, p. 1169–1176.

[19] S. DONNET, A. SAMSON. *Estimation of parameters in incomplete data models defined by dynamical systems*, in "Journal of Statistical Planning and Inference", to appear, 2006.

[20] R. DOUC, A. GUILLIN, J.-M. MARIN, C. ROBERT. *Convergence of adaptive mixtures of importance sampling schemes*, in "Annals of Statistics", to appear, 2007.

[21] M. FROMONT, C. TULEAU. *Lecture Notes in Computer Science (Learning Theory)*, vol. 4005/2006, chap. Functional classification with margin conditions, Springer Berlin / Heidelberg, 2006, p. 94-108.

[22] S. GEY, J.-M. POGGI. *Boosting and Instability for regression trees*, in "Computational Statistics & Data Analysis", vol. 50, 2006, p. 533–550.

[23] F. JAFFRÉZIC, C. MEZA, J.-L. FOULLEY, M. LAVIELLE. *The SAEM algorithm for the analysis of nonlinear traits in genetic studies*, in "Genetics Selection Evolution", to appear, 2006.

[24] W. KENDALL, J.-M. MARIN, C. ROBERT. *Confidence bands for Brownian motion and applications to Monte Carlo simulations*, in "Statistics and Computing", to appear, 2007.

[25] M. LAVIELLE, F. MENTRÉ. *Estimation of population pharmacokinetic parameters of saquinavir in HIV patients and covariate analysis with the SAEM algorithm implemented in MONOLIX*, in "Journal of Pharmacokinetics and Pharmacodynamics", to appear, 2007.

[26] M. LAVIELLE, C. MEZA. *A Parameter Expansion version of the SAEM algorithm*, in "Statistics and Computing", to appear, 2007.

[27] M. LAVIELLE, G. TEYSSIÈRE. *Detection of multiple change–points in multivariae time-series*, in "Lithuanian Mathematical Journal", vol. 46, n^o 4, 2006.

[28] M. LAVIELLE, G. TEYSSIÈRE. *Long-Memory in Economic*, chap. Adaptive Detection of Multiple Change-Points in Asset Price Volatility, Springer-Verlag, 2006.

[29] D. MAKOWSKI, M. LAVIELLE. *Using SAEM to estimate parameters of models of response to applied fertilizer*, in "Journal of Agricultural, Biological, and Environmental Statistics", vol. 11, n^o 1, 2006, p. 45–60.

[30] J.-M. MARIN. *Estimation of variance components for a linear Toeplitz model*, in "Communication in Statistics: Theory and Methods", vol. 36, n^o 12, 2007.

[31] P. MASSART, E. NEDÉLEC. *Risk bounds for statistical learning*, in "Annals of Statistics", vol. 34, n^o 5, 2006.

[32] E. MAZA, J.-M. LOUBES, M. LAVIELLE, L. RODRIGUEZ. *Road trafficking description and short term travel time forecasting, with a classification method*, in "The Canadian Journal of Statistics", vol. 34, n^o 3, 2006, p. 475–491.

[33] J.-M. POGGI, C. TULEAU. *Classification supervisée en grande dimension. Application à l'agrément de conduite automobile*, in "Revue de Statistique Appliquée", vol. LIV, n^o 4, 2006, p. 39–58.

[34] A. SAMSON, M. LAVIELLE, F. MENTRÉ. *Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: application to HIV dynamics model*, in "Computational Statistics & Data Analysis", vol. 51, 2006, p. 1562–1574.

Publications in Conferences and Workshops

[35] A. BAR-HEN, M.-A. POURSAT, P. VANDENKOORNHUYSE. *Influence Function for Phylogenetic Trees*, in "XXIIIrd International Biometric Conference, Montréal", July 2006.

[36] F. BILLY, N. BOUSQUET, G. CELEUX, F. JOSSE. *Vraisemblance d'enchaînements causaux : validation d'un explication a priori confrontée au retour d'expérience*, in "$\lambda\mu$ 15", October 2006.

[37] F. BILLY, N. BOUSQUET, G. CELEUX, E. REMY. *Inférence des paramètres d'une loi de Weibull - Approches classique et bayèsienne*, in "$\lambda\mu$ 15", October 2006.

[38] F. BILLY, N. BOUSQUET, G. CELEUX, E. REMY. *Notions et mesures de cohérence bayésienne entre connaissance a priori et données observées*, in "$\lambda\mu$ 15", October 2006.

[39] N. BOUSQUET, G. CELEUX. *Measures of Bayesian discrepancy between prior beliefs and data knowledge*, in "ESREL 2006", September 2006.

[40] G. CELEUX. *Mixture models for classification (Invited Conference)*, in "Proccedings of GFKS 2006, Berlin", 2006.

[41] N. CHÈZE, J.-M. POGGI. *Outlier detection by iterated boosting*, in "Proceedings of the IFCS'06, Data Science and Classification", Springer, July 2006, p. 213–221.

[42] S. DONNET. *Estimation paramétrique de modèles physiologiques en IRMf*, in "Journées MAS SMAI, Lille", 2006.

[43] R. FRANÇOIS, G. CELEUX, LEFEBVRE, M. LAVIELLE. *Résolution d'un problème inverse par SAEM*, in "38éme Journées de Statistique", May 2006.

[44] M. FROMONT, C. TULEAU. *Les k-plus proches-voisins pour des données fonctionnelles*, in "38éme Journées de Statistique", May 2006.

[45] M. LAVARDE, P. PAMPHILE. *Performances de la sélection de modèles par pénalisation appliquée aux tests accélérés*, in "$\lambda\mu$ 15", October 2006.

[46] M. LAVIELLE, C. LUDENA. *Sélection de modèles linéaires par seuillage aléatoire*, in "Journées MAS SMAI, Lille", 2006.

[47] J.-M. MARIN. *Are risk averse agents more optimistic? A Bayesian estimation approach*, in "he 33rd Seminar of the European Group of Risk and Insurance Economists, Barcelone", 2006.

[48] J.-M. MARIN. *Minimum variance importance sampling via Population Monte Carlo*, in "AMaMef Workshop, Numerical Methods in Finance", 2006.

[49] J.-M. MARIN. *Représentations probabilistes de la méthode des k-plus-proches voisins*, in "Journées MAS SMAI, Lille", 2006.

[50] C. MAUGIS. *Variable selection procedure for transcriptome data clustering*, in "Digiteo Lab Research Group", 2006.

[51] B. MICHEL. *Oil Production: A probabilistic model of Hubbert's curve*, in "Fifth International Conference of the Association for the Study of Peak Oil and Gas (ASPO-5), Pisa, Italy", July 2006.

[52] M. SAUVÉ. *Sélection de modèles de fonctions polynomiales par morceaux en régression*, in "Journées MAS SMAI, Lille", 2006.

[53] N. VERZELEN. *Sélection de voisinage pour des champs de Markov gaussiens*, in "Journées MAS SMAI, Lille", 2006.

[54] L. ZWALD. *Performances statistiques d'algorithmes d'apprentissage : "Kernel Projection Machine" et analyse en composantes principales à noyau*, in "Journées MAS SMAI, Lille", 2006.

Internal Reports

[55] M. AMINGHAFARI, J.-M. POGGI. *Forecasting time series using wavelets*, Technical report, n^{o} 28, Prépublication Université d'Orsay, 2006.

[56] A. BAR-HEN, M.-A. POURSAT, P. VANDENKOORNHUYSE. *Influence function for Phylogenetic reconstruction, a new tool to investigate datasets*, Technical report, submitted, 2006.

[57] S. BENMANSOUR, E. JOUINI, J.-M. MARIN, C. NAPP, C. ROBERT. *Are risk averse agents more optimistic? A Bayesian estimation approach*, Technical report, submitted, 2006.

[58] N. BOUSQUET. *Subjective Bayesian statistics : agreement between prior and data*, Technical report, n^o RR-5900, Institut National de Recherche en Informatique et Automatique, 2006, https://hal.inria.fr/inria-00071367.

[59] G. CELEUX, J.-B. DURAND. *Selecting Hidden Markov Chain States number with Crossvalidated Likelihood*, Technical report, n^o RR-5877, Institut National de Recherche en Informatique et Automatique, 2006, https://hal.inria.fr/inria-00071392.

[60] S. DONNET, A. SAMSON. *Parametric inference for diffusion processes from discrete-time and noisy observations*, Technical report, n^o RR-5809, Institut National de Recherche en Informatique et Automatique, 2006, https://hal.inria.fr/inria-00070215.

[61] P. DRUILHET, J.-M. MARIN. *Equivariant HPD credible sets and MAP estimators*, Technical report, n^o RR-5921, Institut National de Recherche en Informatique et Automatique, 2006, https://hal.inria.fr/inria-00077906.

[62] J.-M. MARIN. *Compatible prior distributions between two nested models from the exponential family*, Technical report, submitted, 2006.

[63] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Analyse en ondelettes de la consommation électrique en vue de l'agrégation et de la désagrégation de courbes pour la prévision de la consommation*, Technical report, n^o December 2005, Rapport de contrat de recherche EDF (84 pages), 2006.

[64] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Classification par ondelettes de courbes pour la prévision de la consommation*, Technical report, n^o June 2006, Rapport de contrat de recherche EDF (88 pages), 2006.

[65] A. SAMSON, M. LAVIELLE, F. MENTRÉ. *The SAEM algorithm for group comparison tests in longitudinal data analysis based on nonlinear mixed-effects model*, Technical report, submitted, 2006.

[66] M. SAUVÉ. *Histogram selection in non gaussian regression*, Technical report, n^o RR-5911, Institut National de Recherche en Informatique et Automatique, 2006, https://hal.inria.fr/inria-00071351.

[67] M. SAUVÉ, C. TULEAU. *Variable selection through CART*, Technical report, n^o RR-5912, Institut National de Recherche en Informatique et Automatique, 2006, https://hal.inria.fr/inria-00071350.