INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team select

Model Selection and Statistical Learning

Futurs

THEME COG

Activity Report

2007

Table of contents

1. Team **1**

2. Overall Objectives **1**

2.1. Model selection in Statistics 1

2.2. Highlights of the year 2

3. Scientific Foundations **2**

3.1. General presentation 2

3.2. A non asymptotic view for model selection 2

3.3. Taking into account the modelling purpose in model selection 2

3.4. Bayesian model selection 2

3.5. Nonlinear mixed effect models 3

4. Application Domains **3**

4.1. General presentation 3

4.2. Curves classification 3

4.3. Reliability 3

4.4. Phylogeny 4

4.5. Population genetics 4

4.6. Neuroimaging 4

4.7. Population pharmacology 4

4.8. Computer Experiments 5

5. Software **5**

5.1. MIXMOD software 5

5.2. MONOLIX software 5

6. New Results **6**

6.1. Model selection in Regression and Classification 6

6.2. Model selection for high-dimensional graphical models 7

6.3. Tests and model selection by resampling 7

6.4. Statistical learning methodology and theory 8

6.5. Adaptive importance sampling schemes 9

6.6. Reliability and Computer Experiments 9

6.7. Classification in genetics 10

6.8. Curves classification, denoising and forecasting 10

6.9. Bayesian estimation 11

6.10. Neuroimaging, Statistical analysis of fMRI data 11

6.11. Nonlinear mixed effects model 11

7. Contracts and Grants with Industry **12**

7.1. Contracts with EDF 12

7.2. Pharmaceutical companies 12

7.3. Other contracts 12

8. Other Grants and Activities **12**

8.1. National Actions 12

8.2. European actions 13

9. Dissemination **13**

9.1. Scientific Communauty animation 13

9.1.1. Editorial responsibilities 13

9.1.2. Invited conferences 13

9.1.3. Scientific animation 13

9.1.4. Invited academics 14

9.2. Teaching 14

10. Bibliography **14**

1. Team

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

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

Administrative assistant
Katia Evrat [TR partially, from July 2007]
Gina Grisvard [TR partially, until June 2007]

Staff member Inria
Marc Lavielle [DR INRIA detached from Université Paris 5, from September 2007, HdR]
Jean-Michel Marin [CR INRIA]

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

Staff member Université Paris 5
Sophie Donnet [ATER Université Paris 5]
Jean-Michel Poggi [Professor Université Paris 5, HdR]

Ph. D. student
Sylvain Arlot [MESR grant]
Pierre Barbillon [MESR grant, from September 2007]
Jean-Patrick Baudry [MESR grant]
Pierre Connault [CIFRE grant, from September 2007]
Robin Genuer [MESR grant]
Merlin Keller [CEA-INRIA grant]
Marc Lavarde [CIFRE grant]
Cathy Maugis [MESR grant]
Bertrand Michel [CIFRE grant]
Vincent Michel [INRIA grant]
Vincent Vandewalle [MESR grant]
Nicolas Verzelen [MESR grant]

Post-doctoral fellow
Nicolas Bousquet [until June 2007]
Lionel Cucala [from September 2007]
Agnès Grimaud

Associate Engineers
Kaelig Chatel [from May 2007]
Anwulin Echenim
Franck Nasse [until Avril 2007]

2. Overall Objectives

2.1. Model selection in Statistics

The research domain for the SELECT project is statistics. Statistical methodology has made great progress over the past few decades, with a variety of statistical learning software packages that support many different methods and algorithms. Users now face the problem of choosing among them, to select the most appropriate method for their data sets and objectives. The problem of model selection is an important but difficult problem both theoretically and practically. Classical model selection criteria, which use penalized minimum-contrast criteria with fixed penalties, are often based on unrealistic assumptions.

SELECT aims to provide efficient model selection criteria with data-driven penalty terms. In this context, SELECT expects to improve the toolkit of statistical model selection criteria from both theoretical and practical perspectives. 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.

2.2. Highlights of the year

Pascal Massart received in June 2007 the Pierre Simon de Laplace prize awarded every four years by the French Statistical Society to a senior statistician.

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 depend on the sample size which make the asymptotic analysis breakdown. An important aim of SELECT is to propose model selection criteria which take 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 that are 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 data-driven penalty choice 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 general ideas that we want to develop, including heuristics derived from the Gaussian theory, special strategies for variable selection and using resampling methods [10], [2].

3.3. Taking into account the modelling purpose in model selection

Choosing a model is not only difficult theoretically. From a practical point of view, it is important to design model selection criteria that accommodate situations in which the data probability distribution P is unknown and which take the model user's purpose into account. Most standard model selection criteria assume that P belongs to one of a set of models, without considering the purpose of the model. By also considering the model user's purpose, we avoid or overcome certain theoretical difficulties and can produce flexible model selection criteria with data-driven penalties. The latter is useful in supervised Classification and hidden-structure models.

3.4. Bayesian model selection

The Bayesian approach to statistical problems is fundamentally probabilistic. A joint probability distribution is used to describe the relationships among all the unknowns and the data. Inference is then based on the posterior distribution i.e. 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 [1], [6], [42], [57] and the calculation of various posterior distributions [1], [6], [40], [42], [61].

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 [17], [18], [28], [24], [25], [15].

4. Application Domains

4.1. General presentation

A key goal of SELECT is to produce methodological contributions in statistics. For this reason, the SELECT team works with applications that serve as an important source of interesting practical problems and require innovative methodologies to address them. Most of our applications involve contracts with industrial partners, e.g. in reliability and pharmacology, although we also have several more academic collaborations, e.g. phylogeny.

4.2. Curves classification

The field of classification for complex data as curves, functions, spectra and time series is important in situations when the values of the explanatory variables of each value are functional, rather than scalar. Classic data analysis questions are being revisited to define new strategies that take the functional nature of the data into account. This new domain, functional data analysis, addresses a variety of applied problems, including longitudinal studies, analysis of fMRI data and spectral calibration.

We are focusing on classification problems with a particular emphasis on clustering, i.e. unsupervised classification ([3], [50]). In addition to classic questions such as the choice of the number of clusters, the norm for measuring the distance between two observations, and the vectors for representing clusters, we must also address a major computational problem. The functional nature of the data requires a very large computational effort, which need to be addressed with efficient or anytime algorithms.

4.3. Reliability

An important theme for SELECT is the problem of aging modelling (or modelling aging), which is funded via a contract with EDF-DER *Fiabilité des Composants et Structures* group. Most French nuclear plants are almost forty years old, the age at which they are no longer warranted to run well. EDF is examining how best to extend the use of nuclear material components beyond forty years and is collaborating with SELECT to analyse the durability of nuclear components.

The other major theme involves changes in reliability processes, based on a contract with Altis. Over the past five years, Altis has drastically changed its chip production process, so that today, half of production involves brass rather than aluminum connections. The previous reliability model is now irrelevant, with abrupt changes in reliability behavior, and SELECT is working on a better model to fit the 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 behavior along the evolutionary tree according to two hidden states, active (ON) or nonactive (OFF). In this research, we are interested in comparing 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 scenarios, 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 allows for 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 combines 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.

In 2007, SELECT started a collaboration with researchers of INRA-SGQA and partners of MIXMOD software on the classification of animal populations using multilocus genotype data.

4.6. Neuroimaging

In 2007, SELECT initiated a working group with team Neurospin (CEA-INSERM-INRIA) on Classification, Statistics and fMRI (functional Magnetic Resonance Imaging) analysis. In this framework two theses are co-supervised by SELECT and Neurospin researchers (Merlin Keller from October 2006 and Vincent Michel from October 2007). 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) and pharmacodynamic (PD) studies (studies investigating the dose-concentration and concentration-effect relationships of drugs) show, for many drugs, a large variability of pharmacokinetic and pharmacodynamic parameters between individuals. Pharmacokinetic parameters describe processes such as absorption, diffusion and metabolism of drugs. The so-called "population PK/PD approach" has been developed to characterize and quantify this variability. 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 ODEs, 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.

4.8. Computer Experiments

In 2007, SELECT initiated several computer experiments, in the framework of conventions with Dassault Aviation and EDF. They concern the resolution of inverses problems using simulation tools to analyse incertainty in highly complex physical systems.

5. Software

5.1. MIXMOD software

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

Participants: Gilles Celeux [Correspondant], Anwulin Echenim.

MIXMOD is being developed in collaboration 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. MIXMOD uses a large variety of algorithms to estimate mixture parameters, e.g., EM, Classification EM, and Stochastic EM. They can be combined to create different strategies that lead to a sensible maximum of the likelihood (or completed likelihood) function. Moreover, different information criteria for choosing a parsimonious model, e.g. the number of mixture component, 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.

At the end of 2006, an expert engineer Anwuli Echenim was hired to improve MIXMOD's performance which is already one of the most complete and rapid sofware on mixture analysis. The last version of MIXMOD includes specific graphical tools to display the results of mixture analysis with qualitative data. The work now consists essentially of continuying the efforts to improve the software performance and to include new Gaussian mixture models specific to the treatement of high dimension data sets.

5.2. MONOLIX software

Keywords: *Non linear mixed effects models*, *SAEM*, *maximum likelihood estimation*.

Participants: Marc Lavielle [Correspondant], Kaelig Chatel, Franck Nasse.

MONOLIX (http://software.monolix.org) is free software dedicated to the analysis of non linear mixed effects models. The objective of the MONOLIX software is to perform:

- Parameter estimation (computing the maximum likelihood estimator of the parameters, without any approximation of the model, computing standard errors for the maximum likelihood estimator),
- Model selection (comparing several models using some information criteria (AIC, BIC), testing hypotheses using the Likelihood Ratio Test, testing parameters using the Wald Test),
- Goodness of fit plots,
- Data simulation.

Several stochastic algorithms are used in MONOLIX: Stochastic approximation of EM (SAEM), Importance Sampling, MCMC, and Simulated Annealing... Theoretical properties of the proposed algorithms and practical applications were published in several papers.

Version 2.2 of MONOLIX is supported by Johnson & Johnson Pharmaceutical Research & Development. Marc Lavielle has presented the software in several occasions:

- PAGANZ meeting, Singapore, February 2007,
- PAGE meeting, Copenhagen, June 2007,
- Roche , Bâle, September 2007,
- TIBOTEC , Mechelen, October 2007,
- Novartis , Bâle, November 2007.

We have obtained from INRIA FUTURS an ODL (Opération Développment Logiciel) to hire engineers (Kaelig Chatel, Franck Nasse). 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, Robin Genuer, Jean-Michel Marin, Pascal Massart, Cathy Maugis, Bertrand Michel, Jean-Michel Poggi, Marie Sauvé, Vincent Vandewalle.

In collaboration with Marie-Laure Martin (INRA), Gilles Celeux and Cathy Maugis [62] developed a variable selection procedure for model-based clustering. The problem is regarded as a model selection problem in the model-based cluster analysis context. A general model generalizing the model of Raftery and Dean (2006) is proposed to specify the role of each variable. This model does not need any prior assumptions about the link between the selected and discarded variables. Models are compared with BIC. The variable's role is obtained through an algorithm embedding two backward stepwise variable selection algorithms for clustering and linear regression. The consistency of the resulting criterion is proved under regularity conditions. The interest of the proposed variable selection procedure is highlighted with numerical experiments on simulated datasets and a genomics application. This last application is the result of a collaboration with researchers of URGV (Evry Genopole). The variable selection procedure is used to extract groups of coexpressed *Arabidopsis thaliana* genes. It allows to improve the clustering and make easier the biological interpretation.

Moreover, the probation period of Fatou Dia (Université Joseph Fourier, Grenoble) leads to a more rapid version of the variable selection procedure written in C++.

Cathy Maugis and Bertrand Michel [48] consider specific Gaussian mixtures to solve simultaneously variable selection and clustering problems. They proposed a non asymptotic penalized criterion to choose the number of mixture components and the relevant variable subset. Because of the non linearity of the associated Kullback-Leibler contrast on Gaussian mixtures, a general model selection theorem for MLE proposed by Massart [2] is used to obtain the penalty function form and the associated oracle inequality. This theorem requires controlling the bracketing entropy of mixture families. Nevertheless, these theoretical results depend on unknown constants. In practice, a "slope heuristic" method ([2]) is applied to calibrate these constants. This joint work is motivated by two practical problems: clustering of transcriptome data [47], [62] and curve classification applied on oil production [49].

Jean-Patrick Baudry, Gilles Celeux and Jean-Michel Marin continued studying the theoretical properties of ICL. They highlight the difference between the notions of components and clusters. The minimum contrast framework offers a new approach to this issue. A new contrast is defined, which is the completed likelihood conditionally to the observations, [32], [33]. It can be justified that it has to do with our classification objective. Applying the slope heuristic of Massart to this new contrast leads to a criterion whose form and behaviour seems to be analogous to those of ICL. ICL should then be considered as a penalised completed likelihood conditionally to the observations criterion, instead of a penalised likelihood criterion whose penalty would take care of the entropy. However, lots of theoretical results are now to be precised. Moreover the difference between the notion of mixture component and cluster has started to be exploited from a data analysis point of view by Jean-Patrick Baudry and Gilles Celeux in collaboration with Adrian Raftery (University of Washington, invited in May) and Naisyin Wang (Texas A & M University).

In collaboration with Christian Robert (Université Paris Dauphine), Gilles Celeux and Jean-Michel Marin have considered Bayesian variable selection in linear regression [42]. They focused on the noninformative case. They shown that, if a Zellner weakly informative prior is used, the model posterior probabilities are sensitive to the choice of an hyperparameter. Consequently, they proposed a new Zellner hierarchical prior. The use of this prior is shown to outperform penalized likelihood criteria and regularisation methods, like the Lasso, in an explicative and a predictive point of view.

In collaboration with Professor Abdallah Mkhadri (University of Marrakesh, Marocco, invited in January), Gilles Celeux and Jean-Michel Marin supervised the thesis of Mohammed El Anbari which concern regularisation methods in linear regression. This year, Mohammed El Anbari has written an exhaustive state of the art on the subject. Comparison with variable selection procedures will be the next step of this study.

Sylvain Arlot and Pascal Massart [4] studied the so-called slope heuristics in the framework of regression on a random design, with possible heteroscedastic noise. Assuming that all the models are made of histograms, they show the same relationship between a “minimal penalty” and an optimal one. This can for instance be used for tuning a penalty, when the optimal penalty is known up to some multiplicative constant. In addition, theoretical and numerical evidence show that penalties linear in the dimension of the models can be suboptimal in some heteroscedastic frameworks. It highlights the interest of considering general shapes of penalties. In general, this optimal shape can be estimated by V-fold or resampling penalties.

Jean-Michel Poggi is the supervisor of the PhD Thesis of Robin Genuer since September 2007 dedicated to Random Forests and related algorithms for variable selection in regression or classification.

6.2. Model selection for high-dimensional graphical models

Participants: Sophie Donnet, Jean-Michel Marin, Pascal Massart, Nicolas Verzelen.

The last decade has witnessed the apparition of applied problems typified by very high-dimensional variables (in marketing database or gene expression studies for instance). Graphical models enable concise representations of associational relations between variables. If the graph is known, the parameters of the model are easily estimated. However, a quite challenging issue is the selection of the most appropriate graph for a given data set.

Pascal Massart and Nicolas Verzelen introduced a penalized criterion based on Mallows heuristic in order to select the graph when the underlying process is known to be stationary. This procedure is proved to satisfy a non asymptotic oracle inequality.

Sylvie Huet (INRA), Pascal Massart, Nicolas Verzelen, and Fanny Villers (INRA) defined a goodness-of-fit test of linear hypotheses for Gaussian regression with Gaussian covariates. They deduced from it a test for Gaussian graphical models. Contrary to most of the existing tests it applies in a high dimensional setting. Besides, it is shown to be minimax against various alternatives. They now carry out numerical experiments with microarray genetic data to assess the graph of genetic networks.

Sophie Donnet and Jean-Michel Marin consider Gaussian graphical models. Bayesian analysis with MCMC methods have been suggested to search over the very high dimensional model space of graphs. In this context, the choice of the hyperparameters of the model is important. Sophie Donnet and Jean-Michel Marin propose an empirical Bayesian procedure combining a MCMC algorithm with a new proposal distribution and a hyperparameters estimation by the SAEM algorithm.

6.3. Tests and model selection by resampling

Participants: Sylvain Arlot, Pascal Massart.

Sylvain Arlot, Gilles Blanchard and Étienne Roquain [31], [30] studied generalized bootstrapped confidence regions for the mean of a random vector whose coordinates have an unknown dependence structure, with a non-asymptotic control of the confidence level. The random vector is supposed to be either Gaussian or to have a symmetric bounded distribution. They consider two approaches, the first one based on a concentration principle and the second one on a direct boostrapped quantile. These results are applied in the one-sided and two-sided multiple testing problem, in which they derive several resampling-based step-down procedures providing a non asymptotic FWER control. According to a simulation study, these procedures can outperform Bonferroni's or Holm's procedures as soon as the observed vector has sufficiently correlated coordinates.

Sylvain Arlot studied model selection by resampling [54], [4], [29]. The classical V-fold cross-validation being biased, a penalization approach is proposed as an alternative. It can be used in a very general framework, and needs the same computation time as V-fold cross-validation. In the case example of regression on histograms, the V-fold penalties lead to a non asymptotic oracle inequality, with constant almost one. This results holds with mild assumptions on the noise level, showing that V-fold penalties are adaptive to heteroscedastic noises. Moreover, a simulation study shows that overpenalization may improve the quality of a model selection procedure, when the sample size is small, as compared to the noise level. The V-fold penalties allowing to choose separately V and the overpenalization factor, they are more flexible than V-fold cross-validation, and outperform it.

Sylvain Arlot and Pascal Massart generalized V-fold penalties to a wide class of resampling penalties [54], [4], [29]. In the histogram regression case, a non asymptotic oracle inequality and adaptation to the smoothness of the regression function and the heteroscedastic noise are proven. In the classification case, some first steps towards an oracle inequality are proven [4]. This shows that resampling penalties may be an efficient alternative to local Rademacher complexities. A simulation study in regression shows that resampling penalties outperform classical procedures such as Mallows' C_p and V-fold cross-validation.

6.4. Statistical learning methodology and theory

Participants: Gilles Celeux, Jean-Michel Marin, Pascal Massart, Vincent Vandewalle.

The k-nearest-neighbour procedure is a well-known method used in supervised classification. While it has been superseded by more recent methods developed in machine learning, it remains an essential tool for classifiers. In collaboration with Christian Robert (Université Paris Dauphine) and Mike Titterington (University of Glasgow, Scotland), Jean-Michel Marin proposes a reassessment of this approach as a statistical technique derived from a proper probabilistic model [61], [40]; in particular, they modify the assessment made in a previous analysis of this method where the underlying probabilistic model was not completely well-defined. Once clear probabilistic bases of the k-nearest-neighbour procedure are established, they proceed to the derivation of practical computational tools to conduct Bayesian inference on the parameters of the corresponding model. In particular, they assess difficulties inherent in pseudo-likelihood and in path sampling approximations of a missing normalising constant, and propose a perfect sampling strategy to implement a correct MCMC sampler associated with our model. Illustrations of the performance of the corresponding Bayesian classifier are provided for two benchmark datasets, demonstrating in particular the limitations of the pseudo-likelihood approximation in this set-up.

In collaboration with Christophe Biernacki (Université de Lille) and Gérard Govaert (UTC Compiègne), Gilles Celeux continues, in a Bayesian framework, to study a predictive approach for clustering for the latent class model to analyse multivariate multinomial discrete data sets. From the algorithmic point of view it appaers that Gibbs sampling algorithms often outperforms the evolutionary algorithms developed with Marc Schoenauer and Damien Tessier from TAO team (INRIA Futurs). From the statistical point of view, this fully non-informative approach appears, as expected, to outperform for small sample size its asymptotic approximation, the ICL criterion. From the data analysis point of view, the features of the predictive clustering remain mysterious. Moreover, it seems it could be unattractive in some situations.

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. This year, they investigated the appliability of model selection criteria in the semi-supervised context. At first, they clarify the role of different focus. They characterise situations where unlabeled data can degrade classification accuracy. In this purpose they developed, in a Bayesian framework, two specific model selection criteria. The first criterion proposes weighting unlabeled data to limit their influence. The second criterion puts into question the simultaneous use of labeled and unlabeled data.

6.5. Adaptive importance sampling schemes

Participant: Jean-Michel Marin.

There are some different strategies of *adaptive importance sampling*. For instance, the Population Monte Carlo scheme can be implemented as the D-kernel algorithm ([19], [20]), which tries to find the best mixture of D given kernels or importance sampling distributions in terms of either minimum variance or minimum Kullback-Leibler divergence. While this algorithm is shown to converge to the best possible solution, it is within a small family of distributions and thus may fail to represent properly the target distribution. In collaboration with Christian Robert (Université Paris Daupine) and Antonietta Mira (University of Varese, Italy), Jean-Michel Marin studied a different perspective to merge together different importance samples [41]. The proposed adaptive algorithm does not require any tuning parameter. In collaboration with Jean-Marie Cornuet (INRA, Montepellier), this scheme has been tested on population genetics models.

In collaboration with Olivier Cappé (École Nationale des Télécommunications, Paris), Randal Douc (École Polytechnique, France), Arnaud Guillin (École Centrale Marseille) and Christian Robert, Jean-Michel Marin studied the performances of a simultaneous update of both weights and parameters in a mixture of transition kernels when used in a Population Monte Carlo environment and geared towards the minimization of an entropy criterion [59]. Convergence to the optimal solution is established and the performances of the proposed scheme are studied on artificial and real examples.

In collaboration with Roberto Casarin (University of Brescia, Italy), Jean-Michel Marin compare three regularized particle filters in an online data processing context [60]. They carried out the comparison in terms of hidden states filtering and parameter estimation, considering a Bayesian paradigm and a univariate stochastic volatility model. They discussed the use of an improper prior distribution in the initialization of the filtering procedure and showed that the Regularized Auxiliary Particle Filter outperforms the Regularized Sequential Importance Sampling and the Regularized Sampling Importance Resampling.

6.6. Reliability and Computer Experiments

Participants: Pierre Barbillon, Nicolas Bousquet, Gilles Celeux, Agnès Grimaud, Marc Lavarde, Pascal Massart, Jean-Michel Marin.

In the framework of a contrat with EDF concerning reliability, Nicolas Bousquet and Gilles Celeux have studied the behavior of three different discrete failure time models in a highly censored setting for material breaking down at prompting. Two models, an extension to the discrete setting of the Weibull model an a model related to a Polya urn scheme show good behavior. However, for this last model, maximum likelihood estimation results in poor performance in this highly censored data setting and informative Bayesian analysis is desirable. They are now studying possible reasonable prior distributions to get an honest Bayesian estimation procedure.

In the framework of a convention with EDF, Gilles Celeux and Agnès Grimaud worked in collaboration with Yannick Lefebvre and Étienne de Rocquigny on the resolution of not linear inverse problems for the quantification of uncertainties in a physical model. More precisely, noisy observed data (Y) were dependent, through a known but complex and expensive function H from non-observed data X. The aim is to estimate parameters of the probability distribution of the non observed data (X) and the variance of the noise. The problem has a missing data structure and can be solved with an EM-type algorithm.

For the first step, a linear approximation was considered about a fixed vector x_0. A simple characterisation of the identifiabilty of the model was exhibited, after which the EM algorithm and accelerated version the ECME algorithm were used to estimate the parameters [37]. They now work with the true complex fonction H. In order to avoid too many calls to the expensive function H, they proposed to couple the Stochastic EM algorithm with a kriging step where H is not computed exactly at each time. An alternative Importance Sampling algorithm was also developed by Yannick Lefebvre.

Many scientific phenomena are now investigated by complex models or code. A computer experiments consists of a number of runs of the code with various inputs. In general, the output of a computer experiments is deterministic (rerunning the code with the same inputs gives identical observations). The aim of computer experiments is to fit a predictor of the output to the data. In this paradigm, Jean-Michel Marin is the supervisor of the PhD thesis of Pierre Barbillon since September 2007 [55]. The goal of this thesis is to construct adaptive experimental design using Importance Sampling methodology.

6.7. 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 account for the existence of missing data and some a priori biological information. Moreover, to classify the genes they make use of the variable selection for mixture models developed in [62].

6.8. Curves classification, denoising and forecasting

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

In collaboration with Christine Tuleau (Université Paris-Sud and then Université Nice), Jean-Michel Poggi proposed a classification of complex data for objectivization [51]. This real world problem consists of explaining the subjective drivability using physical criteria coming from signals measured during experiments. They suggest an approach for the discriminant variables selection trying to take advantage of the functional nature of the data. The problem is ill-posed, since the number of explanatory variables is significantly greater than the sample size. The strategy proceeds in three steps: first, signal pre-processing, including wavelet denoising and synchronization, second dimensionality reduction by compression using a common wavelet basis, and finally, the selection of useful variables using a stepwise strategy involving successive applications of the CART method.

In collaboration with Michel Misiti (Ecole Centrale de Lyon), Yves Misiti (Université Paris-Sud), G. Oppenheim (Université Marne la Vallée), Jean-Michel Poggi proposed a wavelet-based procedure for clustering signals [50]. It combines an individual signal preprocessing by wavelet denoising, a dimensionality reduction step by wavelet compression and a classical clustering strategy applied to a suitably chosen set of wavelet coefficients. The ability of wavelets to cope with signals of arbitrary or time dependent regularity as well as to concentrate signal energy in few large coefficients, offers a useful tool to carry out both significant noise reduction and efficient compression. A simulated example and an electrical data set are considered to illustrate the value of introducing wavelets for clustering such complex data.

In collaboration with Michel Misiti (Ecole Centrale de Lyon), Yves Misiti (Université Paris-Sud), G. Oppenheim (Université Marne la Vallée), Jean-Michel Poggi published a book entitled “Wavelets and their applications” [3]. The last fifteen years have seen an explosion of interest in wavelets with applications in fields such as image compression, turbulence, human vision, radar and earthquake prediction. Wavelets represent an area that combines signal in image processing, mathematics, physics and electrical engineering. This book is intended for the wide audience that is interested in mastering the basic techniques, such as decomposition and compression.

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 [49] proposed a probabilistic model of oil production in a homogeneous geological zone.

6.9. Bayesian estimation

Participant: Jean-Michel Marin.

In collaboration with Pierre Druilhet (ENSAI, Rennes), Jean-Michel Marin [21] 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 Christian Robert (Université Paris Dauphine), Jean-Michel Marin published a book entitled “Bayesian Core: A Practical Approach to Computational Bayesian Statistics” [1].

6.10. Neuroimaging, Statistical analysis of fMRI data

Participants: Gilles Celeux, Sophie Donnet, Robin Genuer, Merlin Keller, Christine Kéribin, Marc Lavielle, Jean-Michel Marin, Vincent Michel, Jean-Michel Poggi.

Tow new thesis projects have begun, as part of a collaboration with Neurospin
(http://www.math.u-psud.fr/select/reunions/neurospin/Welcome.html).
Vincent Michel's thesis, started in October 2007, addresses supervised Classification of fMRI images.

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 described as the non-analytical solution of a differential system. The input of this model are neuronal efficiencies. In order to test variability in the neuronal efficiencies, Sophie Donnet and Marc Lavielle proposed a method combining a general method to estimate the parameters for regression models defined by differential system –developed by Sophie Donnet and Adeline Samson [18] and a statistical test. A first study on a real data set extracted from the primary visual cortex has proved that the more flexible model –including neuronal efficiencies variability– is better than the usual model. A non-linearity of the neuronal activity has also been detected. To confirm these conclusions, a study on more individuals, more regions of interests is in progress with Jean-Baptiste Poline.

Merlin Keller began his PhD in October 2006 under the supervision of Alexis Roche (CEA, Neurospin) and Marc Lavielle. Merlin Keller and Alexis Roche developped a method for fMRI group analysis, which goes beyond the standard approach in that it accounts for measurement errors which can vary from one subject to another, and possible non-normal distribution of the effect within the population under study [64], [39]. Simulations as well as experiments on real fMRI datasets showed that this method has the potential to enhance the detection of activated brain regions.

6.11. Nonlinear mixed effects model

Participants: Sophie Donnet, Marc Lavielle, Adeline Samson.

The MONOLIX group (http://software.monolix.org), 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 [17], [18], [28], [24], [25], [15].

7. Contracts and Grants with Industry

7.1. Contracts with EDF

Participants: Nicolas Bousquet, Gilles Celeux, Agnès Grimaud, Jean-Michel Marin.

- SELECT has a contract with EDF regarding discrete failure models.
- SELECT has a contrat with EDF regarding modelling uncertainty in deterministic models.
- SELECT has a contrat with EDF regarding classification of flawed electric transformer.

7.2. Pharmaceutical companies

Participant: Marc Lavielle.

- SELECT is negotiating a contract with Pfizer and a contract with Tibotec.

7.3. Other contracts

Participants: Pierre Connault, Marc Lavarde, Pascal Massart, Bertrand Michel.

- 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.
- SELECT has a contract with Meyssier-Dowty (CIFRE grant of Pierre Connault) on the problem of variable selection in high dimension.

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 aggregation methods. Most of SELECT members are involved in this working group.

SELECT is animating a working group on Classification, Statistics and fMRI imaging with Neurospin.

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 various 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). This multi-disciplinary group, born in October 2003, has been meeting every month.

Moreover, Marc Lavielle is responsible of an ANR project (projet blanc) on the MONOLIX software which started in 2006.

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 was invited two weeks in the Department of Economics of Varese University. He gave a conference during his stay.

9. Dissemination

9.1. Scientific Communauty animation

9.1.1. Editorial responsibilities

Participants: Gilles Celeux, Pascal Massart, Jean-Michel Poggi.

- Gilles Celeux is Editor-in-Chief of *Statistics and Computing*. He is Associate Editor of *Journal de la SFdS*, *CSBIGS* and *La Revue Modulad*.
- Pascal Massart is Associated Editor of *Annales de l'IHP*, *Annals of Statistics*, *Journal de la SFdS*, *ESAIM Proceedings* and *Foundations and Trends in Machine Learning*.
- Jean-Michel Poggi is Associated Editor of *Journal de la SFdS* and *CSBIGS*.

9.1.2. Invited conferences

Participants: Gilles Celeux, Jean-Michel Marin, Pascal Massart, Marc Lavielle.

- Gilles Celeux was invited speaker at the Workshop on Model- Based Clustering in Dublin (July 2007) and at a workshop on finite mixture analysis in Leuwen (November 2007).
- Gilles Celeux was one of the organisers and contributors of an International Workshop in Lisieux on Statistical approaches and validation in clustering (June 2007).
- Jean-Michel Marin was invited speaker at the sixth workshop on Objective Bayes Methodology in Rome (June 2007).
- Jean-Michel Marin was invited speaker at Spring Bayes 2007 meeting in Australia (September 2007). During that period, Jean-Michel Marin was invited two weeks in the School of Mathematical Sciences of Queensland University of Technology (Australia).
- Pascal Massart was invited speaker at the annual meeting of the French Statistical Society in Angers (June 2007).
- Marc Lavielle was invited speaker and gave a MONOLIX course at the PAGANZ 2007 meeting (Singapore, February 2007).
- Marc Lavielle gave a plenary talk at the PAGE 2007 meeting (Copenhagen, June 2007).

9.1.3. Scientific animation

Participants: Gilles Celeux, Jean-Michel Marin, Pascal Massart, Marc Lavielle, Jean-Michel Poggi.

- Gilles Celeux has chaired the evaluation council of Jouy-en-Josas unit of MIA (Mathématiques et Informatique Appliquées) Department of INRA. He is a member of the scientific council of the MIA Department of INRA. He was member of the evaluation coucil of the Department EPFA (Écologie des Forêts, Prairies et Milieux Aquatiques) of INRA.
- Marc Lavielle is the director of the GDR (Groupement de Recherche) "Statistique et Santé", unité de recherche 3067 of the CNRS.
- Marc Lavielle is a member of the council of the SMAI (Société de Mathématiques Appliquées et Industrielles).

- Marc Lavielle is a member of the scientific council of the CIMPA (Centre Inernational de Mathématiques Pures et Appliquées).
- Jean-Michel Marin is the head of the council of the French Statistical Society.
- Pascal Massart is the head of the Department of Mathematics of University Paris-Sud.
- Pascal Massart is a member of the scientific council of Euradom and of the working group on "le rôle des mathématiques dans le monde contemporain" of the *Académie des Sciences*.
- Jean-Michel Poggi is a member of the council of the French Statistical Society (SFdS).
- Jean-Michel Poggi is a member of the CNU 26 (Conseil National des Universités).

9.1.4. Invited academics

- Adrian Raftery (University of Washington, USA) during one month in May.
- Abdallah Mkhadri (University of Marrakesh, Marocco) in January for two weeks.
- Rik Lopuhäa (University of Delft, Netherlands) in July for two weeks.
- Roberto Casarin (University of Brescia, Italia) in July for two weeks.

9.2. Teaching

Pascal Massart is responsible of the M2 "Modélisation stochastique et statistique" of University Paris-Sud. 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, Berlin / Heidelberg, 2007.

[3] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Wavelets and their applications*, ISTE Publishing Knowledge, 2007.

Doctoral dissertations and Habilitation theses

[4] S. ARLOT. *Ré-échantillonnage et sélection de modèles*, Ph. D. Thesis, Université Paris-Sud, 2007.

[5] M. LAVARDE. *Fiabilité des semi-conducteurs, tests accélérés, Sélection de modèles définis par morceaux et application à la détection de sur-stress*, Ph. D. Thesis, Université Paris-Sud, 2007.

[6] J.-M. MARIN. *Habilitation à Diriger des Recherches : Méthodes de Monte-Carlo adaptatives et statistique bayésienne*, Ph. D. Thesis, Université Paris Dauphine, 2007.

Articles in refereed journals and book chapters

[7] M. AMINGHAFARI, J.-M. POGGI. *Forecasting time series using wavelets*, in "International Journal of Wavelets, Multiresolution and Information Processing", vol. 5, n^o 5, 2007, p. 709–724.

[8] S. BenMansour, E. Jouini, J.-M. Marin, C. Napp, C. Robert. *Are risk averse agents more optimistic ? A Bayesian estimation approach*, in "Journal of Applied Econometrics", to appear, 2007.

[9] C. Biernacki, G. Celeux, A. Echenim, G. Govaert, F. Langrognet. *Le logiciel MIXMOD d'analyse de mélange pour la classification et l'analyse discriminante*, in "La Revue Modulad", vol. 35, 2007, p. 25-44.

[10] L. Birgé, P. Massart. *Minimal penalties for Gaussian model selection*, in "Probability Theory and Related Fields", vol. 138, 2007, p. 33–73.

[11] G. Blanchard, O. Bousquet, P. Massart. *Statistical performance of support vector machines*, in "Annals of Statistics", to appear, 2007.

[12] N. Bousquet. *Avantages et enjeux de l'analyse statistique bayésienne en durée de vie industrielle*, in "Lettre Techniques de l'Ingénieur : Risques Industriels", vol. 23, n^{o} 2, 2007.

[13] G. Celeux. *Mixture Models for Classification*, in "Studies in Classification, Data Analysis, and Knowledge Organization", Springer, Berlin / Heidelberg, 2007, p. 3-14.

[14] G. Celeux, J.-B. Durand. *Selecting Hidden Markov Chain States Number with Cross-Validated Likelihood*, in "Computational Statistics", to appear, 2007.

[15] E. Comets, C. Verstuyft, M. Lavielle, P. Jaillon, L. Becquemont, F. Mentré. *Modelling the influence of MDR1 polymorphism on digoxin pharmacokinetic parameters*, in "European Journal of Clinical Pharmacology", vol. 63, 2007, p. 437–449.

[16] G. Consonni, J.-M. Marin. *Mean field variational Bayesian inference for latent variable models*, in "Computational Statistics & Data Analysis", vol. 52, n^{o} 2, 2007, p. 790–798.

[17] S. Donnet, A. Samson. *Estimation of parameters in incomplete data models defined by dynamical systems*, in "Journal of Statistical Planning and Inference", vol. 50, 2007, p. 2381–2398.

[18] S. Donnet, A. Samson. *Parametric inference for mixed models defined by stochastic differential equations*, in "ESAIM Probability and Statistics", to appear, 2007.

[19] R. Douc, A. Guillin, J.-M. Marin, C. Robert. *Convergence of adaptive mixtures of importance sampling schemes*, in "Annals of Statistics", vol. 35, n^{o} 1, 2007, p. 420–448.

[20] R. Douc, A. Guillin, J.-M. Marin, C. Robert. *Minimum variance importance sampling via Population Monte Carlo*, in "ESAIM: Probability and Statistics", vol. 11, 2007, p. 427–447.

[21] P. Druilhet, J.-M. Marin. *Invariant HPD credible sets and MAP estimators*, in "Bayesian Analysis", to appear, 2007.

[22] R. François, J.-M. Marin. *Initiation à R*, in "La revue MODULAD", to appear, 2007.

[23] W. Kendall, J.-M. Marin, C. Robert. *Confidence bands for Brownian motion and applications to Monte Carlo simulations*, in "Statistics and Computing", vol. 17, n^{o} 1, 2007, p. 1–10.

[24] 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", vol. 34, n^{o} 2, 2007, p. 229–249.

[25] M. LAVIELLE, C. MEZA. *A Parameter Expansion version of the SAEM algorithm*, in "Statistics and Computing", vol. 17, n^{o} 2, 2007, p. 121–130.

[26] M. LAVIELLE, C. LUDEÑA. *Random thresholds for linear model selection*, in "ESAIM Probability and Statistics", to appear, 2007.

[27] S. ROBIN, S. SCHBATH, V. VANDEWALLE. *Statistical tests to compare motif count exceptionalities*, in "BMC Bioinformatics", vol. 8, n^{o} 1, 2007, 84.

[28] A. SAMSON, M. LAVIELLE, F. MENTRÉ. *The SAEM algorithm for group comparison tests in longitudinal data analysis based on nonlinear mixed-effects model*, in "Statistics in Medicine", to appear, 2007.

Publications in Conferences and Workshops

[29] S. ARLOT. *Sélection de modèles par ré-échantillonnage*, in "Congrès National de la SMAI, Praz-sur-Arly", June 2007.

[30] S. ARLOT, G. BLANCHARD, E. ROQUAIN. *Resampling-based confidence regions and multiple tests for a correlated random vector*, in "Pascal workshop and Pascal challenge “Type I and type II errors for Multiple Simultaneous Hypothesis Testing”, Paris", May 2007.

[31] S. ARLOT, G. BLANCHARD, E. ROQUAIN. *Resampling-based confidence regions for the mean*, in "Lecture Notes in Artificial Intelligence 4539, COLT 2007, Berlin", 2007, p. 127–141.

[32] J. BAUDRY. *Clustering Through Model Selection Criteria*, in "Workshop on Statistical approaches and validation in clustering: Mixture models and nonparametric methods, Lisieux", June 2007.

[33] J. BAUDRY. *Critères de selection de modèles pour la classification non supervisée*, in "Deuxièmes Rencontres des Jeunes Statisticiens, Aussois", September 2007.

[34] G. CELEUX. *Predictive Clustering*, in "Workshop on Model-Based Clustering, Dublin", July 2007.

[35] G. CELEUX, G. GOVAERT. *Clustering mixture models: general aspects and model selection criteria*, in "Workshop on Statistical approaches and validation in clustering: Mixture models and nonparametric methods, Lisieux", June 2007.

[36] G. GOVAERT, G. CELEUX. *Some specific aspects of clustering mixture models*, in "Workshop on Statistical approaches and validation in clustering: Mixture models and nonparametric methods, Lisieux", June 2007.

[37] A. GRIMAUD. *Résolution de problèmes inverses non linéaires á l'aide de l'algorithme EM et de la méthode Circé*, in "Groupe de Travail Méthodes inverses en probabilité et statistiques, EDF, Clamart", December 2006.

[38] A. GRIMAUD. *Résolution de problèmes inverses non linéaires á l'aide des algorithmes EM et ECME*, in "Séminaire des thèses du département MRI, EDF, Chatou", Mars 2007.

[39] M. KELLER, A. ROCHE, S. MÉRIAUX. *A mixed-effect statistic for two-sample group analysis in fMRI*, in "Thirteenth annual meeting of the Organization Human Brain Mapping", 2007.

[40] J.-M. MARIN. *A Bayesian reassessment of nearest-neighbour classification*, in "Spring Bayes 2007, Coolangatta, Australia", September 2007.

[41] J.-M. MARIN. *Adaptive Multiple Importance Sampling*, in "Workshop on Bioinformatics, Genetics and Stochastic Computation: Bridging the Gap, Banff International Research Station, Canada", July 2007.

[42] J.-M. MARIN. *Variable selection in Gaussian linear regression*, in "The sixth International Workshop on Objective Bayesian Analysis, University La Sapienza, Rome, Italy", June 2007.

[43] C. MAUGIS, G. CELEUX, M.-L. MARTIN-MAGNIETTE. *Sélection de variables pour la classification par mélange gaussiens*, in "39ème Journées de Statistique, Angers", June 2007.

[44] C. MAUGIS. *Sélection de variables pour la classification par mélanges gaussiens*, in "39ème Journées de Statistique, Angers", June 2007.

[45] C. MAUGIS. *Variable selection for clustering with Gaussian mixture models*, in "Statistique Mathématique et Applications, CIRM, Marseille", 2007.

[46] C. MAUGIS. *Variable selection for transcriptome data clustering with Gaussian mixture models*, in "2ème Rencontres des Jeunes Statisticiens, Aussois", September 2007.

[47] C. MAUGIS. *Variable selection for transcriptome data clustering with Gaussian mixture models*, in "Fifth Workshop of Statistical methods for post-genomic data, Paris", January 2007.

[48] C. MAUGIS, B. MICHEL. *A Penalized Criterion for Gaussian Mixture Model Selection. A variable selection and clustering problem*, in "International Meeting on Empirical Processes and Asymptotic Statistics, Rennes", June 2007.

[49] B. MICHEL. *Modélisation de la production d'hydrocarbure au sein d'un bassin pétrolier*, in "39ème Journées de Statistique, Angers", June 2007.

[50] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Clustering signals using wavelets*, in "Computational and ambient intelligence: 9th International Work-Conference on Artificial Neural Networks, IWANN 2007", Springer, Lecture Notes in Computer Science, 4507, 2007.

[51] J.-M. POGGI, C. TULEAU. *Classification of objectivization data using CART and wavelets*, in "Proceedings of the IASC 07, Aveiro, Portugal", June 2007.

[52] D. TESSIER, M. SCHOENAUER, C. BIERNACKI, G. CELEUX, G. GOVAERT. *Parameter Setting for Evolutionary Latent Class Clustering*, in "Advances in Computation and Intelligence, Second International Symposium, ISICA 2007", Springer, Lectures Notes in Computer Science, 4683, 2007, p. 472–484.

[53] V. VANDEWALLE. *Model selection with semi-supervised data: a classification focus*, in "Deuxièmes Rencontres des Jeunes Statisticiens, Aussois", September 2007.

Internal Reports

[54] S. Arlot. *Model selection by resampling penalization*, Technical report, Arxiv:math.ST/0701542, 2007, http://arxiv.org/abs/math.ST/0701542.

[55] P. Barbillon. *Modèles réduits à partir d'expériences numériques*, Technical report, Mémoire de Master Recherche, Université Paris-Sud, 2007.

[56] N. Bousquet. *A Bayesian analysis of industrial lifetime data with Weibull distributions*, Technical report, n° RR-6025, Institut National de Recherche en Informatique et Automatique, 2007, http://hal.inria.fr/inria-00115528.

[57] N. Bousquet. *Diagnostics of prior-data conflict in applied Bayesian analysis*, Technical report, n° RR-5900, Institut National de Recherche en Informatique et Automatique, 2007, http://hal.inria.fr/inria-00071367.

[58] N. Bousquet, G. Celeux. *Étude de modèles de défaillance à la sollicitation en fiabilité industrielle*, Technical report, Rapport de contrat de recherche EDF-INRIA (June), 2007.

[59] O. Cappé, R. Douc, A. Guillin, J.-M. Marin, C. Robert. *Adaptive Importance Sampling in General Mixture Classes*, Technical report, n° RR-6332, Institut National de Recherche en Informatique et Automatique, 2007, http://hal.inria.fr/inria-00181474.

[60] R. Casarin, J.-M. Marin. *Online data processing: comparison of Bayesian regularized particle filters*, Technical report, n° RR-6153, Institut National de Recherche en Informatique et Automatique, 2007, http://hal.inria.fr/inria-00138007.

[61] J.-M. Marin, C. Robert, D. Titterington. *A Bayesian reassessment of nearest-neighbour classification*, Technical report, n° RR-6173, Institut National de Recherche en Informatique et Automatique, 2007, http://hal.inria.fr/inria-00143783.

[62] C. Maugis, G. Celeux, M.-L. Martin-Magniette. *Variable selection for Clustering with Gaussian Mixture Models*, Technical report, n° RR-6211, Institut National de Recherche en Informatique et Automatique, 2007, http://hal.inria.fr/inria-00153057.

[63] M. Misiti, Y. Misiti, G. Oppenheim, J.-M. Poggi. *Optimisation, pilotée par la prévisibilité, de partitions pour la prévision par désagrégation de la courbe de charge*, Technical report, Rapport de contrat de recherche EDF (55 pages), 2007.

[64] A. Roche, S. Mériaux, M. Keller, B. Thirion. *Mixed-effect statistics for group analysis in fMRI: a general maximum likelihood approach*, Technical report, CEA, 2007.