INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team select

Model Selection and Statistical Learning

Saclay - Île-de-France

THEME COG

Table of contents

1. Team

Research Scientist

Gilles Celeux [Team Vice-Leader, DR INRIA]
Marc Lavielle [DR INRIA detached from Université Paris 5]
Jean-Michel Marin [CR INRIA until August 2008]

Faculty Member

Pascal Massart [Team Leader, Professor Université Paris-Sud]
Christine Keribin [Assistant Professor]
Marie-Anne Poursat [Assistant Professor]
Jean-Michel Poggi [Professor Université Paris 5]

Technical Staff

Kaelig Chatel
Anwulin Echenim [until June 2008]
Jean-François Si Abdallah [from November 2008]

PhD Student

Pierre Barbillon [MESR grant]
Jean-Patrick Baudry [MESR grant]
Pierre Connault [CIFRE grant]
Mohammed El Anbari [France-Marocco grant]
Robin Genuer [MESR grant]
Merlin Keller [CEA-INRIA grant]
Cathy Maugis [MESR grant]
Bertrand Michel [CIFRE grant]
Vincent Michel [INRIA grant]
Vincent Vandewalle [MESR grant]
Nicolas Verzelen [MESR grant]

Post-Doctoral Fellow

Sylvain Arlot [ATER until August 2008]
Lionel Cucala [until August 2008]
Agnès Grimaud [until August 2008]

Administrative Assistant

Katia Evrat [TR partially]

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

Sylvain Arlot has been hired as CR by CNRS and Jean-Michel Marin has been hired as Professor by Université of Montpellier in September 2008.

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.

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 and the calculation of various posterior distributions.

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.

4. Application Domains

4.1. Introduction

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. 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 concerns fatigue rupture analysis based on a research collaboration based on a research collaboration with SAFRAN an high-technology group (Aerospace propulsion, Aicraft equipment, Defense Security, Communications). The aim is to perform an efficient statistical control of aircraft equipment production processes.

The LASSO is a selection method for linear regression. It minimizes the sum of squared errors, with a penalty on the sum of the absolute values of the coefficients. The objectif of Pierre Connault, in his PhD, is to calibrate automatically that penalty.

The fatigue strength of aeronotics equipments is tested from cyclic applications of stress on small steel test-tubes. How do these results on test-tubes can be extrapolated to reliability analysis of the entire equipment ? A probabilistic model is proposed by SELECT.

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 2008, SELECT continue a collaboration with researchers of INRA-SGQA on the classification of animal populations using multilocus genotype data through the course of Vincent Benezech (ENS Ulm).

4.6. Neuroimaging

Since 2007 SELECT participates to 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 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.

4.8. Environment

This year, a study has been achieved by Jean-Michel Poggi, François-Xavier Jollois (Université Paris-Descartes) and Bruno Portier (INSA de Rouen), in the context of a collaboration between AirNormand, Paris Descartes University and INSA of Rouen,. They analyzed PM10 pollution during 2004-2006 in Rouen area using six different monitoring sites to quantify the effects of variables of different types, mainly meteorological versus other pollutant measurements. Three recent non parametric statistical methods (random forests, mixture of linear models and nonlinear additive models) have been used and beyond the application, this study shed light on those methods [49], [67], [68]

4.9. Computer Experiments

Since 2007, SELECT developed several computer experiment studies, 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, Jean-François Si Abdallah.

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.

Fron July 2006 to June 2008 an expert engineer, Anwuli Echenim, worked to improve MIXMOD's performance. And MIXMOD is 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. Moreover, new Gaussian mixture models specific to the treatement of high dimension data sets have been included. Since November 2008, a new expert engineer Jean-François Si Abdallah has been hired for two years to continue to enrich the software, improve the peerformances, code a proper graphical library for clustering displays in MIXMOD and propose a version available via internet.

In December 2008, the second meeting of the MIXMOD users has been organized in Lille.

5.2. MONOLIX software

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

Participants: Marc Lavielle [Correspondant], Kaelig Chatel.

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.

Marc Lavielle has presented the software in several occasions:

- Novartis , Cambridge & New-Jersey, May 2008.
- PAGE meeting, Marseille, June 2008,
- J&J, Beerse, July 2008,
- Sanofi-Aventis, Chilly-Mazarin, October 2008,
- GSK, London, October 2008,
- Roche , Bâle, December 2008,

Version 2.4 of MONOLIX is available since October 2008. This version of the software was supported by Johnson & Johnson Pharmaceutical Research & Development.

The MONOLIX Project consists primarily in developing the next versions of the MONOLIX software with a view to raising its level of functionalities and responding to major requirements of the bio-pharmaceutical industry.

The MONOLIX Project is carried out by INRIA, and sponsored by the Industry.

The MONOLIX Scientific Guidance Committee involves representatives of the sponsors.

We have obtained from INRIA Saclay-Île-de-France an ADT (Action Développment Logiciel) to hire two engineers (Kaelig Chatel, Hector Mesa which will come in 2009).

6. New Results

6.1. Model selection in Regression and Classification

Participants: Sylvain Arlot, Jean-Patrick Baudry, Gilles Celeux, Lionel Cucala, Mohammed El Anbari, Robin Genuer, Jean-Michel Marin, Pascal Massart, Cathy Maugis, Bertrand Michel, Jean-Michel Poggi, Vincent Vandewalle.

In collaboration with Marie-Laure Martin-Magniette (INRA), Gilles Celeux and Cathy Maugis [21] 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. They proposed a model generalizing the model of Raftery and Dean (2006) 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. Variable role is obtained through an algorithm embedding two backward stepwise variable selection algorithms for clustering and linear regression. The model identifiability is established and 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. The DNA microarray technology generating many missing values, an extension of the variable selection procedure taken the existence of missing entries into account is proposed. It avoids the missing entry imputation usually used in preprocessing. Currently, they are interested in an improvement of the variable role modelling. This modelling consisting of partitioning the irrelevant variables according to their dependence or independence with some relevant clustering variables, is suggested to avoid an overpenalization of some models [69].

Cathy Maugis and Bertrand Michel consider specific Gaussian mixtures to solve simultaneously variable selection and clustering problems. In [70], 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 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 [71], they study the practical use of their penalized criterion. A "slope heuristic" method is applied to calibrate these constants. This joint work is motivated by two practical problems: clustering of transcriptome data and curve classification applied on oil production [53], [2]. Moreover, Jean-Patrick Baudry, Cathy Maugis and Bertrand Michel began a practical study of the application of the slope heuristic and of the solutions to the practical difficulties it involves.

Sylvain Arlot and Pascal Massart [7], [41] 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 general, the optimal shape of the penalty can be estimated by V-fold or resampling penalties.

Jean-Patrick Baudry, Gilles Celeux and Jean-Michel Marin deepened the study of the estimator and the mixture model selection procedures they introduced in the clustering framework, and particularly for the choice of the number of clusters to be designed. On the one hand, they continued the simulation work about it, aiming especially at understanding their asymptotic behavior. On the other hand, they continued trying to get theoretical results about those criteria. The theoretical framework has been precised and the properties to set are well identified [54], [42].

Selecting the numbers of components in a mixture model has been studied a lot in the last years by the members of SELECT. However, doing the same for a model containing spatial dependence is a new and challenging problem. Lionel Cucala and Jean-Michel Marin study such problems with applications in pattern recognition. The spatial dependence is brought by the introduction of a Potts model, a classical Markov random field model. The application to pattern recognition consists of discriminating the pixels of an image obtained by tomography. They show that a modified version of the ICL criterion gives excellent results.

In collaboration with Professor Abdallah Mkhadri (University of Marrakesh, Marocco), 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 proposed a new regularisation method. This method can be view as a variant of the method Elastic Net of Zou and Hastie (2005)[1] taking account of the correlation

between predictors. He is now working on extensive numerical experiments to compare all the regularisation methods proposed in the 2000 year for regression. This comparison includes Bayesian variable selection methods in linear regression.

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. Random Forest, due to Leo Breiman in 2001, proceeds by aggregation decision trees according to two random perturbations. The first one perturbs the learning sample according to the bootstrap principle and the second one acts on the covariate space by choosing randomly a small number of explanatory variables to split a tree node. Surprisingly, this algorithm is extremely powerful for regression and classification problems, not only for prediction but also for variable selection purposes. The PhD thesis is articulated following three directions:

- The preliminary theoretical direction concerns mathematical understanding of the reasons of this amazing behaviour.
- The second methodological direction aims at improving the knowledge about how to tune the parameters. It includes computer intensive simulations and comparisons based on well-known real data sets [65], [46].
- The last one is of applied nature and takes place on the joint working group between SELECT and Neurospin (INRIA, CEA) dedicated to statistical methods for fMRI new data in order to improve knowledge about brain activities. It aims to develop ad-hoc variable selection strategies.

6.2. Bayesian model selection

Participant: Jean-Michel Marin.

Gibbs random fields are polymorphous statistical models that can be used to analyse different types of dependence, in particular for spatially correlated data. However, when those models are faced with the challenge of selecting a dependence structure, the use of standard model choice methods is hampered by the unavailability of the normalising constant in the Gibbs likelihood. In particular, from a Bayesian perspective, the computation of the posterior probabilities of the models under competition requires special likelihood-free simulation techniques like the Approximate Bayesian Computation (ABC) algorithm that is intensively used in population Genetics. In collaboration with Christian Robert (Université Paris Dauphine), Aude Grelaud (INRA Jouy-en-Josas), François Rodolphe (INRA Jouy-en-Josas) and Jean-François Taly (INRA Jouy-en-Josas), Jean-Michel Marin shows how to implement an ABC algorithm geared towards model choice in the general setting of Gibbs random fields, demonstrating in particular that there exists a sufficient statistic across models [66]. The accuracy of the approximation to the posterior probabilities can be further improved by importance sampling on the distribution of the models. The practical aspects of the method are detailed through two applications, the test of an independent Bernoulli model versus a first-order Markov chain, and the choice of a folding structure for a protein of *Thermotoga maritima* implicated into signal transduction processes.

In Scott (2002)[2] and Congdon (2006)[3], a new method is advanced to compute posterior probabilities of models. It is based solely on MCMC outputs restricted to single models, bypassing reversible jump and other model exploration techniques. However, Jean-Michel Marin in collaboration with Christian Robert (Université Paris-Dauphine) show [22], [36] that the proposals of Scott and Congdon are biased and advance several arguments towards this thesis, the primary one being the confusion between model-based posteriors and joint pseudo-posteriors.

6.3. Selection of high dimensional graphical models

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

[1] Journal of the Royal Statistical Society, B
[2] Journal of the American Statistical Association
[3] Computational Statistics and Data Analysis

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.

Sylvie Huet (INRA), Pascal Massart, Nicolas Verzelen, and Fanny Villers (INRA) [23] 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 have also carried out numerical experiments with microarray genetic data and have assessed the graph of genetic networks [24].

Nicolas Verzelen [73] also considered the problem of estimation in a linear regression setting with Gaussian covariates. He introduced a novel model selection method that deals with high-dimensional data. This procedure is proved to satisfy a non asymptotic oracle inequality and minimax adaptive properties. Contrary to other procedures, the rates of convergence do not depend on the correlation between the covariates. Nicolas Verzelen, Christophe Giraud (École Polytechnique), and Sylvie Huet (INRA) now apply this technique for estimating the graph of a Gaussian graphical model.

Noel Cressie (Ohio State University) and Nicolas Verzelen [14] introduced a method for approximating the distribution of a stationary Gaussian field by a Gaussian graphical model. This technique enables to speed up Markov Chain Monte Carlo algorithms and more generally Bayesian techniques. Pascal Massart and Nicolas Verzelen [3] are currently working on a model selection procedure that performs both estimation and approximation.

In collaboration with Sophie Donnet (Université Paris-Dauphine), 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 [35].

6.4. Tests and model selection by resampling

Participants: Sylvain Arlot, Pascal Massart.

Sylvain Arlot studied model selection by V-fold cross-validation or penalization [56], [32], [29], [31], [30], [27]. 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 generalized V-fold penalties to a wide class of resampling penalties [57]. 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. A simulation study in regression shows that resampling penalties outperform classical procedures such as Mallows' C_p and V-fold cross-validation.

In collaboration with Gilles Blanchard and Étienne Roquain, Sylvain Arlot [5] 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 we 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.

6.5. Statistical learning methodology and theory

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

In collaboration with Peter L. Bartlett, Sylvain Arlot [58] studied adaptivity to the margin condition in statistical learning, in the context of model selection. A classical condition for fast learning rates is the margin condition, first introduced by Mammen and Tsybakov. They considered a weaker version of this condition that allows to take into account that learning within a small model can be much easier than in a large one. Requiring this "strong margin adaptivity" makes the model selection problem more challenging. They first proved, in a general framework, that some penalization procedures (including local Rademacher complexities) exhibit this adaptivity when the models are nested. Contrary to previous results, this holds with penalties that only depend on the data. Second, they proved that strong margin adaptivity is not always possible when the models are not nested: for every model selection procedure (even a randomized one), a problem exists for which the procedure does not demonstrate strong margin adaptivity.

Model-based clustering consists of fitting a Gaussian mixture model to data and identifying each cluster with one of its components. In practice, however, individual clusters can be poorly fitted by Gaussian distributions, and in that case model-based clustering tends to represent one non-Gaussian cluster by a mixture of two or more Gaussian distributions. Jean-Patrick Baudry and Gilles Celeux propose first selecting the total number of Gaussian mixture components, K, using BIC and then combining them hierarchically according to an entropy criterion related to the ICL criterion. This yields a unique soft clustering for each number of clusters less than or equal to K; The interest of the method has been highlight with flow cytometry data. This research has been initiated during the visit of Adrian Raftery (University of Washington to SELECT in 2007 [60].

In collaboration with Christophe Biernacki (Université de Lille) and Gérard Govaert (UTC Compiègne), Gilles Celeux propose non asymptotic version of integrated likelihood for the latent class model or multivariate multinomial mixture model. They exploit the fact that a fully Bayesian analysis with Jeffreys non informative prior distributions does not involve technical difficulty to propose an exact expression of the integrated *complete-data* likelihood, which is known as being a meaningful model selection criterion in a clustering perspective. Similarly, they propose a Monte Carlo approximation of the integrated *observed-data* likelihood via a Bayesian importance sampling strategy. Those exact and the approximate criteria favorably compete respectively with their standard asymptotic BIC approximations for choosing the number of mixture components. This research gives the opportunity to highlight the deep purpose difference between the integrated *complete-data* and the *observed-data* likelihoods: The integrated *complete-data* likelihood is focussing on a cluster analysis view and favors well separated clusters, implying some robustness against model misspecification, while the integrated *observed-data* likelihood is focussing on a density estimation view and is expected to provide a consistent estimation of the distribution of the data [62].

Vincent Vandewalle pursues his PhD thesis about semi-supervised model-based classification under the supervision of his advisors Christophe Biernacki (Université de Lille), Gilles Celeux and Gérard Govaert(UTC). He focused on the discriminant analysis situation which is of main interest for applications. Firstly, he explored the possibility of discarding or not unlabeled data to learn the classification rule through hypothesis testing [40]. Then, he is investigating specific information based criteria for model selection in the semi-supervised setting. It seems to be a promising strategy since some significant improvements of standard criteria (AIC, BIC) have been obtained in the related context of supervised classification (BEC criterion of Bouchard and Celeux (2006)[4]).

[4] IEEE on PAMI

Jean-Michel Poggi proposed a procedure for detecting outliers in regression problems. It is based on information provided by boosting regression trees. The key idea is to select the most frequently resampled observation along the boosting iterations and reiterate after removing it. The selection criterion is based on Tchebychev's inequality applied to the maximum over the boosting iterations of the average number of appearances in bootstrap samples. Thus, the procedure is noise distribution free. A lot of well-known bench data sets are considered and a comparative study against two well-known competitors allows to show the interest of the method [25].

6.6. Adaptive importance sampling schemes

Participant: Jean-Michel Marin.

Population Monte Carlo has been introduced as a sequential importance sampling technique to overcome poor fit of the importance function. In collaboration with Alessadra Iacobucci and Christian Robert (Université Paris-Dauphine), Jean-Michel Marin compare the performances of the original Population Monte Carlo algorithm with a modified version that eliminates the influence of the transition particle via a double Rao-Blackwellisation. This modification is shown to improve the exploration of the modes through a large simulation experiment on posterior distributions of mean mixtures of distributions.

Sequential techniques can be added to the Approximate Bayesian Computation (ABC) algorithm to enhance its efficiency. Sisson et al. (2007)[5] introduced the ABC-PRC algorithm to improve upon existing ABC-MCMC algorithms. In collaboration with Marc Beaumont (University of Reading), Jean-Marie Cornuet (Imperial College) and Christian Robert, Jean-Michel Marin show that, while the ABC-PRC method is based upon the theoretical developments of Del Moral et al. (2006)[6], the application to the ABC setting induces a bias in the approximation to the posterior distribution of interest [61]. It is however possible to devise an alternative version based on genuine importance sampling arguments that they call ABC-PMC in connection with the population Monte Carlo method introduced in Cappé et al. (2004)[7]. This algorithm is simpler than the ABC-PRC algorithm, it does not suffer from the original bias, and it includes an automatic scaling of the forward kernel. Moreover, when applied to a population genetics example, its efficiency compares favourably with two standard ABC algorithms.

In collaboration with Jean-Marie Cornuet, Antonietta Mira (University of Insubria, Varese) and Christian Robert, Jean-Michel Marin study how it is possible to recycle all the past simulations in an adaptive importance sampling scheme. They propose the Adaptive Multiple Importance Sampling (AMIS) algorithm. The AMIS scheme is aimed at an optimal recycling of past simulations in an iterated importance sampling scheme. The difference with earlier adaptive importance sampling implementations like Population Monte Carlo is that the importance weights of all simulated values, past as well as present, are recomputed at each iteration, following the technique of the deterministic multiple mixture estimator of Owen and Zhou (2000)[8]. Although convergence properties of the algorithm are difficult to be fully investigated, they demonstrate through a population genetics example that the improvement brought by this technique is significant.

6.7. Reliability and Computer Experiments

Participants: Pierre Barbillon, Gilles Celeux, Pierre Connault, Agnès Grimaud, Pascal Massart, Jean-Michel Marin.

In the framework of a convention with EDF, Gilles Celeux and Agnès Grimaud worked in collaboration with Yannick Lefebvre and Étienne de Rocquigny (EDF) 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.

[5] Proc. Natl. Acad. Sci.USA
[6] Journal of The Royal Statistical Society B
[7] Journal of Computational and Graphical Statistics
[8] Journal of the American Statistical Association

In 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. This year to reduce the influence of the linearisation point they proposed a solution using an iterative linearisation of the function H. Then, with Pierre Barbillon, they develop methods using the exact complex fonction H. In order to avoid too many calls to this expensive function H, they propose a non-linearised method coupling the use of the Stochastic EM algorithm with a MCMC method and a kriging approximation of the H function [47], [63]. This method cmpares favorably with an alternative Importance Sampling algorithm proposed by Yannick Lefebvre.

For an other convention with EDF, Gilles Celeux and Côme Roero student of ENSAI) analyzed the practical impact of various models for the reliability of discrete lifetime data. They show that the Polya urn model is of limited interest since it does not allow for accelerated aging models. Moreover, by numerical experiments, they show that there is little interest to consider alternative discrete Weibull models to the standard Weibull model [64].

In aircraft equipment, fatigue is one of the first cause of ruptures. Morevover fatigue ruptures appear brutally and can be catastrophics: important material damage, human death. The fatigue rupture is a complex random process: it is influenced by numerous and various factors of the production process and environment. The large number of factors, and the strong variability of some of them, yield any expertise very difficult.

SELECT develops a collaboration with SAFRAN via the Phd of Pierre Connault, supervised by Pascal Massart and Patrick Pamphile (Université Paris-Sud). Variable selection methods (CART, LASSO) have been used to perform an efficient statistical control of aircraft equipment production processes. LASSO is a regularisation method for linear regression. It minimizes the sum of squared errors, with a penalty on the sum of the absolute values of the coefficients. The objectif of Pierre Connault is to calibrate automatically that penalty. Moreover, a probabilistic model of fatigue has been proposed to extrapolate results on tets tubes to assess the reliability of the entire equipment.

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. The goal of this thesis is to construct adaptive experimental design using Importance Sampling methodology. In collaboration with Yves Auffray (Université Paris-Sud), Pierre Barbillon and Jean-Michel Marin consider the context of kernel interpolation for which the concept of Native Spaces is central. These spaces are constructed from conditionally positive definite kernels. They proposed to generalize the usual definition of this type of kernels. Based on this new definition, they show that it is possible to construct properly the corresponding Native Spaces. Then, they give the interpolation operators and show that they are the same than the classic ones.

6.8. Classification in genomics

Participants: Gilles Celeux, Cathy Maugis.

Following the Cathy Maugis thesis [1], we decide to use her material in collaboration with biologists of URGV (INRA, Evry Genopole) and Marie-Laure Martin-Magniette (INRA) to improve functional annotation of *Arabidopsis thaliana* genes. This joint work with URGV is expected to enter an ANR proposal SONATA in Spring 2009.

6.9. Curves classification, denoising and forecasting

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

In collaboration with Anestis Antoniadis (Université J. Fourier, Grenoble) and Irène Gijbels (Leuven University), Jean-Michel Poggi considered a non parametric noisy data model $Y_k = f(x_k) + \epsilon_k, \quad k = 1, \cdots, n,$ where the unknown signal f from $[0, 1]$ in $\mathbf{R}$ is assumed to belong to a wide range of function classes, including discontinuous functions and the $\epsilon_k' s$ are independent identically distributed noises with zero median. The unknown distribution of the noise is assumed to have heavy tails, so that no moments of the noise exist. The design points are assumed to be deterministic points, not necessarily equispaced within the interval $[0, 1]$. Standard kernel methods cannot be applied in this situation. Their approach first uses local medians to construct variables Z_k structured as a Gaussian nonparametric regression, then they apply a wavelet block penalizing procedure adapted to non equidistant designs to construct an estimator of the regression function. Under mild assumptions on the design, they show that their estimator, which has a good practical behavior, simultaneously attains the optimal rate of convergence over a wide range of Besov classes, without prior knowledge of the smoothness of the underlying functions or prior knowledge of the error distribution [4].

In order to take into account the variation of EDF (the French electrical company) portfolio due to the liberalization of the electrical market, it is essential to conveniently disaggregate the global signal. The idea is to disaggregate the global load curve in such a way that the sum of disaggregated predictions improve significantly the prediction of the global signal considered as a whole. In collaboration with Michel Misiti (Ecole Centrale de Lyon), Yves Misiti (Université Paris-Sud), G. Oppenheim (Université Marne la ée), Jean-Michel Poggi designs a strategy to optimize with respect to a predictability index, a preliminary clustering of individual load curves. The optimized clustering scheme is directed by forecasting performance via a cross-prediction dissimilarity index and proceeds as a discrete gradient type algorithm [55], [72].

- Forecasting time series using wavelets

In collaboration with Mina Aminghafari (Amirkabir University, Teheran), Jean-Michel Poggi made uses of wavelets in a statistical forecasting purpose for time series. Recent approaches involve wavelet decompositions in order to handle non stationary time series. They study and extended an approach proposed by Renaud et al, to estimate the prediction equation by direct regression of the process on the Haar non-decimated wavelet coefficients depending on its past values. The new variants are used first for stationary data and after for stationary data contaminated by a deterministic trend [37].

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

6.10. Neuroimaging, Statistical analysis of fMRI data

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

This research takes place 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. It is supervised by Gilles Celeux, Christine Keribin and Bertrand Thirion (Parietal). During his first year, he have adressed the question of decoding cognitive information from functional magnetic resonance (MR) images using classification techniques. He studied classification methods (SVM, LDA) and feature selection methods (Anova, Manova, Mutual Information...). Bertrand Thirion and Vincent Michel have developed a multivariate approach based on a mutual information criterion, estimated by nearest neighbors, which can handle a large number of dimensions and is able to detect non linear relations between the features and the labels [39]. Functional data are huge (more than 100,000 voxels by image, for dozens of images), and he is now trying to deal with this limitation with methods, including clustering methods, allowing to reduce the dimensionality of the problem. Moreover, Vincent Michel and Robin Genuer examine the value of random forests to deal with such problems.

Merlin Keller began his PhD in October 2006 under the supervision of Alexis Roche (CEA, Neurospin) and Marc Lavielle. During his second year as a PhD student, He have worked on the problem of activation detection in fMRI group data analysis. In this context, He have investigated the use of an adaptive thresholding method, which estimates the number of activated voxels using model selection techniques. He have also worked on defining regions of interest using an atlas, rather than through a pre-defined activity threshold, as is the classical heuristic. He is also currently working on a multivariate model of fMRI group data which accounts for spatial normalization errors. This continues the work done last year where he proposed a method which accounted for spatial uncertainty, but which was not based on a multivariate model[38], [18].

Christine Keribin has achieved a bibliographic analysis of variational Bayesian methods for spatial mixture models.

6.11. Robust phylogenetic reconstructions

Participants: Christine Keribin, Marie-Anne Poursat.

Maximum likelihood methods in phylogeny are based on an explicit DNA or protein sequence evolution model. Depending on the complexity of the model, the robutness of the inferred tree is to be assessed.

Based on the computation of the influence function, a tool to measure the impact of each piece of sampled data on the statistical inference, we analyze the support of the maximum likelihood tree for each site and provide a new tool for filtering datasets in the context of maximum likelihood phylogenetic reconstructions [8].

The most commonly used test of reliability of an inferred tree is the Felsenstein's bootstrap. Numerous studies explored the advantages and limitations of the bootstrap to evaluate branch support in phylogeny. We are working on an alternative : we propose a test for branches of evolutionary trees, based on an appropriate version of the standard likelihood-ratio statistic.

6.12. Nonlinear mixed effects model

Participant: Marc Lavielle.

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 [19].

7. Contracts and Grants with Industry

7.1. Contracts with EDF

Participants: Gilles Celeux, Agnès Grimaud, Jean-Michel Poggi.

- 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 wavelet analysis of the electrical load consumption for the aggregation and desaggregation of curves to improve total signal prediction.

7.2. Pharmaceutical companies

Participant: Marc Lavielle.

- Pfizer
- Tibotec.

7.3. Other contracts

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

- 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 contrat with SAFRAN - MESSIER-DOWTY, an high-technology group (Aerospace propulsion, Aicraft equipment, Defense Security, Communications),regarding modelling reliability of Aircraft Equipment (collaboration with Patrixk Pamphile (Université Paris-Sud).
- SELECT has a contract) with Total regarding short time Fourier transform for Spurious signal detection.
- SELECT has a collaboration with AirNormand regarding the statistical analysis of PM10 air pollution in the Haute-Normandie area.

7.4. Project GAS

Participants: Gilles Celeux, Pascal Massart, Bertrand Michel.

The project GAS was selected by the DIGITEO consortium in the framework of the "Domaines d'Intérêt Majeur" call of the Région ële-de-France. The main partner is GEOMETRICA. The other partners of the project are the Ecole Polytechnique (F. Nielsen) and SELECT. The project intends to explore and to develop new researches at the crossing of information geometry, computational geometry and statistics. It started in September 2008 and it is expected duration is two years. In this setting, Pascal Massart is the cosupervisor with Frédéric Chazal (GEOMETRICA) of the thesis of Claire Caillerie (GEOMETRICA).

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 regularisation methods in regression. 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.

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 *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: Sylvain Arlot, Gilles Celeux, Jean-Michel Marin, Marc Lavielle.

- Sylvain Arlot was invited speaker at the European Mathematical meeting in Oberwolfach.
- Gilles Celeux was invited speaker at COMPSTAT 2008 in Porto and to the 32th meeting of the German Classification Society.
- Gilles Celeux and Pascal Massart were invited speakers at the "Journée d'Apprentissage" in Nancy in December 2008.
- Marc Lavielle was invited speaker at IBC 2008 (Dublin).
- Jean-Michel Marin was invited speaker at Journées Statistique du Sud in Toulouse (June 2008).
- Jean-Michel Marin was invited speaker at Journées MAS SMAI in Rennes (August 2008).

9.1.3. Scientific animation

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

- Gilles Celeux is 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 director of the GDR (Groupement de Recherche) "Statistique et Santé", Research Unit 3067 of the CNRS.
- Marc Lavielle is member of the council of the SMAI (Société de Mathématiques Appliquées et Industrielles).
- Marc Lavielle is member of the scientific council of the CIMPA (Centre International de Mathématiques Pures et Appliquées).
- Marc Lavielle is member of the Comité de Préfiguration à la Haute Autorité sur les OGM.
- 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.
- Pascal Massart coorganised the international meeting of Mathematical Statistics and applications of Fréjus.
- Jean-Michel Poggi is Cochair seminar of Probabity ans Statistics of the "laboratoire de Mathématiques d'Orsay", seminar ECAIS (Extraction de connaissances : approches informatiques et statistiques) of IUT de Paris 5 Descartes and of "Séminaire Parisien de Statistique".
- Jean-Michel Poggi is member of the Council of the French statistical society (SFdS).
- Jean-Michel Poggi is member of the Board of the "Environment group" of the French statistical society (SFdS).

9.1.4. Invited academics

- Andrew Barron (Yale University) during two weeks in September.
- Terry Speed (University of California at Berkeley) during two weeks in September.

9.2. Teaching

All the SELECT members are teaching in various courses of different universities and in particular in the M2 "Modélisation stochastique et statistique" of University Paris-Sud.

10. Bibliography

Year Publications

Doctoral Dissertations and Habilitation Theses

[1] C. MAUGIS. *Sélection de variables pour la classification non supervisée par mélanges gaussiens. Application à l'étude de données transcriptomes*, Ph. D. Thesis, Université Paris-Sud, 2008.

[2] B. MICHEL. *Modélisation de la production d'hydrocarbures dans un bassin pétrolier*, Ph. D. Thesis, Université Paris-Sud, 2008.

[3] N. VERZELEN. *Modèles graphiques gaussiens et sélection de modèles*, Ph. D. Thesis, Université Paris-Sud, 2008.

Articles in International Peer-Reviewed Journal

[4] A. ANTONIADIS, I. GIJBELS, J.-M. POGGI. *Smoothing non equispaced heavy noisy data with wavelets*, in "Statistica Sinica", to appear, 2008.

[5] S. ARLOT, G. BLANCHARD, É. ROQUAIN. *Some non-asymptotic results on resampling in high dimension, I: confidence regions*, in "Annals of Statistics", To appear, 2008.

[6] S. ARLOT, G. BLANCHARD, É. ROQUAIN. *Some non-asymptotic results on resampling in high dimension, II: multiple tests*, in "Annals of Statistics", To appear, 2008.

[7] S. ARLOT, P. MASSART. *Data-driven calibration of penalties for least-squares regression*, in "Journal of Machine Learning Research", to appear, 2008.

[8] A. BAR-HEN, M. MARIADASSOU, M.-A. POURSAT, P. VANDENKOORNHUYSE. *Influence Function for Robust Phylogenetic Reconstructions*, in "Molecular Biology and Evolution", vol. 25, 2008, p. 869-873.

[9] S. BENMANSOUR, E. JOUINI, J.-M. MARIN, C. NAPP, C. ROBERT. *Are risk agents more optimistic? A Bayesian estimation approach*, in "J. Appl. Econometrics", vol. 23, 2008, p. 843–860.

[10] G. BLANCHARD, O. BOUSQUET, P. MASSART. *Statistical performance of support vector machines*, in "Annals of Statistics", vol. 36, 2008, p. 489-531.

[11] O. CAPPÉ, R. DOUC, A. GUILLIN, J.-M. MARIN, C. ROBERT. *Adaptive importance sampling in general mixture classes*, in "Stat. Comput.", (to appear), 2008.

[12] G. CELEUX, J.-B. DURAND. *Selecting hidden Markov model state number with cross-validated likelihood*, in "Computational Statistics", vol. 23, 2008, p. 541-564.

[13] J.-M. CORNUET, F. SANTOS, M. BEAUMONT, C. ROBERT, J.-M. MARIN, D. BALDING, T. GUILLEMAUD, A. ESTOUB. *Infering population history with DIY ABC: a user-friedly approach Approximate Bayesian Computation*, in "Bioinformatics", (to appear), 2008.

[14] N. CRESSIE, N. VERZELEN. *Conditional-mean least-squares fitting of Gaussian Markov random fields to Gaussian fields*, in "Comput. Statist. and Data Analysis", vol. 52, n^{o} 5, 2008, p. 2794–2807.

[15] L. CUCALA, J.-M. MARIN, C. ROBERT, M. TITTERINGTION. *A Bayesian reassessment of nearest-neighbour classification*, in "J. Amer. Statist. Assoc.", (to appear), 2008.

[16] S. DONNET, A. SAMSON. *Parametric inference for mixed models defined by stochastic differential equations*, in "ESAIM Probability and Statistics", vol. 12, 2008, p. 196–218.

[17] A. IACOBUCCI, J.-M. MARIN, C. ROBERT. *On variance stabilisation by double Rao-Blackwellisation*, in "Comput. Statist. Data Anal.", (to appear), 2008.

[18] M. KELLER, A. ROCHE, B. THIRION. *Dealing with spatial normalization errors in fMRI group inference using hierarchical modeling*, in "Statistica Sinica", vol. 18, 2008, p. 1357-1374.

[19] M. LAVIELLE, C. LUDEÑA. *Random thresholds for linear model selection*, in "ESAIM Probability and Statistics", vol. 12, 2008, p. 173–195.

[20] J.-M. MARIN, C. ROBERT. *Approximating the marginal likelihood in mixture models*, in "Indian Bayesian Society News Letter", vol. V, n^{o} 1, 2008, p. 2–7.

[21] C. MAUGIS, G. CELEUX, M.-L. MARTIN-MAGNIETTE. *Variable selection for Clustering with Gaussian Mixture Models*, in "Biometrics", To appear, 2008.

[22] C. ROBERT, J.-M. MARIN. *On some difficulties with a posterior probability approximation technique*, in "Bayesian Anal.", vol. 3, n^{o} 2, 2008, p. 427–442.

[23] N. VERZELEN, F. VILLERS. *Goodness-of-fit Tests for high-dimensional Gaussian linear models*, in "Annals of Statistics", to appear, 2008.

[24] N. VERZELEN, F. VILLERS. *Tests for Gaussian graphical models*, in "Comput. Statist. Data Analysis", to appear, 2008.

Articles in National Peer-Reviewed Journal

[25] N. CHÈZE, J.-M. POGGI. *Détection de données aberrantes en régression*, in "Revue des Nouvelles Technologies de l'Information", 2008, p. 159-171.

[26] P. MASSART. *Sélection de modèle: de la théorie à la pratique. Journal de la SFDS, 2008.*, in "Journal de la SFdS", to appear, 2008.

Articles in Non Peer-Reviewed Journal

[27] S. ARLOT. *Comment choisir un modèle?*, in "Le Mensuel de l'Université", May 2008.

[28] G. CELEUX. *Book Review: Data Clustering, Theory Algorithms and Applications by Gan, G., Chaoqun, M. A., and Wu, J.*, in "Biometrics", vol. 64, 2008, p. 656-657.

Invited Conferences

[29] S. ARLOT. *V-fold cross-validation improved: V-fold penalization*, in "Journées Statistiques du Sud, Toulouse", June 2008.

[30] S. ARLOT. *V-fold penalization: an alternative to V-fold cross-validation*, in "Cherry Bud Workshop, Keio University, Yokohama, Japan", March 2008.

[31] S. ARLOT. *V-fold penalization: an alternative to V-fold cross-validation*, in "Second Canada-France Congress, Montreal, Canada", March 2008.

[32] S. ARLOT. *V-fold penalization: an alternative to V-fold cross-validation*, in "Oberwolfach Reports. Vol. 4, no. 4, Zürich", Mathematisches Forschungsinstitut Oberwolfach Report, vol. 4, n^{o} 4, European Mathematical Society (EMS), 2008.

[33] G. CELEUX. *Choosing the Number of Clusters in the Latent Class Model*, in "German Classification Society 32nd Annual Conference, Joint Conference with the British Classification Society (BCS) and the Dutch/Flemish Classification Society (VOC), Hamburg", July 2008.

[34] G. CELEUX. *Traitement statistique des petits échantillons*, in "Journées Fiabilité des Matériaux et des Structures, Nantes", March 2008.

[35] J.-M. MARIN. *An empirical Bayes procedure for the selection of Gaussian graphical models*, in "Journées de Statistique du Sud, Toulouse", June 2008.

[36] J.-M. MARIN. *On some computational methods for Bayesian model choice*, in "Journées MAS SMAI, Rennes", August 2008.

International Peer-Reviewed Conference/Proceedings

[37] M. AMINGHAFARI, J.-M. POGGI. *Multistep Wavelet-based Forecasting Time Series*, in "Proceedings, International Symposium of Forecasting 2008, Nice", June 2008.

[38] M. KELLER, S. MÉRIAUX, A. ROCHE, P. PINEL, B. THIRION. *A mixed-effect statistic for two-sample group analysis in fMRI*, in "Biomedical Imaging: From Nano to Macro, 2007. ISBI 2008. 5th IEEE International Symposium on", May 2008.

[39] V. MICHEL, C. DAMON, B. THIRION. *Mutual Information-Based Feature Selection Enhances fMRI brain activity classification*, in "Biomedical Imaging: From Nano to Macro, 2007. ISBI 2008. 5th IEEE International Symposium on", May 2008.

[40] V. VANDEWALLE, C. BIERNACKI, G. CELEUX, G. GOVAERT. *Are unlabeled data useful in semi-supervised model-based classification? combining hypothesis testing and model choice*, in "First joint meeting of the Société Francophone de Classification and the CLassification And Data Analysis Group of SIS, Caserta", June 2008.

Workshops without Proceedings

[41] S. ARLOT, P. MASSART. *Data-driven calibration of penalties for least squares regression*, in "Statistique Mathématique et Applications, Fréjus", September 2008.

[42] J.-P. BAUDRY. *Classification non supervisée par minimisation d'un contraste ad hoc. Calibrage du critère de sélection de modèle correspondant par heuristique de pente.*, in "Groupe de travail INAPG-Select, Paris", March 2008.

[43] J.-P. BAUDRY. *Clustering through Contrast Minimization. Calibration of a Model Selection Criterion by Slope Heuristic.*, in "Statistique Mathématique et Applications, Fréjus", September 2008.

[44] J.-P. BAUDRY. *Clustering Through Model Selection Criteria. Slope Heuristic.*, in "Working Group on Applied, Bayesian and Computational Statistics, Seattle", July 2008.

[45] J.-P. BAUDRY, G. CELEUX, R. GOTTARDO, A. RAFTERY. *Combining Gaussian Components for Clustering. Mixtures of Mixtures.*, in "Congrès conjoint de la Société Statistique du Canada et de la Société Française de Statistique, Ottawa", May 2008.

[46] R. GENUER, J.-M. POGGI, C. TULEAU. *Random Forests: an Experimental Study*, in "Statistique Mathématique et Applications, Fréjus", September 2008.

[47] A. GRIMAUD. *Estimation de la variabilité pour des systèmes multidimensionnels par des méthodes inverses linéaires et non linéaires.*, in "Séminaire de Probabilités et Statistique, Université Paris-Sud, Orsay, Orsay", March 2008.

[48] A. GRIMAUD. *Estimation de la variabilité pour des systèmes multidimensionnels par des méthodes inverses linéaires et non linéaires.*, in "Séminaire de Statistique, Université de Caen, Caen", March 2008.

[49] F.-X. JOLLOIS, J.-M. POGGI, B. PORTIER. *Influences météorologiques et apport des phénomènes â l'échelle régionale : une approche statistique*, in "Colloque "Les particules dans l'air", Rouen", September 2008.

[50] C. MAUGIS. *A data-driven penalized criterion for Gaussian mixture model selection*, in "Congrès conjoint de la SSC et de la SFdS, Ottawa", May 2008.

[51] C. MAUGIS. *A data-driven penalized criterion for Gaussian mixture model selection*, in "Workshop: Parametric and nonparametric mixture model and their applications, Pau", June 2008.

[52] C. MAUGIS. *A non-asymptotic penalized criterion for Gaussian mixture model selection*, in "Statistique Mathématique et Applications, Fréjus", September 2008.

[53] B. MICHEL. *Estimation of the oil exploration process in an hydrocarbon basin*, in "Statistiques mathématiques et applications, Fréjus", September 2008.

Scientific Books (or Scientific Book chapters)

[54] J.-P. BAUDRY, G. CELEUX, J.-M. MARIN. *Selecting models focussing on the modeller purpose*, in "COMPSTAT 2008: Proceedings in Computational Statistics, Heidelberg", Physica, 2008, p. 337-348.

[55] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Optimized Clusters for Disaggregated Electricity Load Forecasting*, in "COMPSTAT 2008: Proceedings in Computational Statistics, Heidelberg", Physica, 2008, p. 225-232.

Research Reports

[56] S. ARLOT. *V-fold cross-validation improved: V-fold penalization*, Technical report, arXiv:0802.0566, 2008, http://fr.arxiv.org/abs/0802.0566.

[57] S. ARLOT. *Model selection by resampling penalization*, Technical report, hal-00262478, 2008, http://hal.archives-ouvertes.fr/hal-00262478/en/.

[58] S. ARLOT, P. BARTLETT. *Margin adaptive model selection in statistical learning*, Technical report, arXiv:0804.2937, 2008, http://fr.arxiv.org/abs/0804.2937.

[59] S. ARLOT, P. MASSART. *Slope heuristics for variable selection and clustering via Gaussian mixtures*, Technical report, n^{o} RR-6556, Institut National de Recherche en Informatique et Automatique, 2008, http://hal.inria.fr/inria-00287631/fr/.

[60] J.-P. BAUDRY, A. RAFTERY, G. CELEUX, K. LO, R. GOTTARDO. *Combining Mixture Components for Clustering*, Technical report, n^{o} 6644, Institut National de Recherche en Informatique et Automatique, 2008, http://hal.inria.fr/inria-00321090/fr/.

[61] M. BEAUMONT, J.-M. CORNUET, J.-M. MARIN, C. ROBERT. *Adaptivity for ABC algorithms: the ABC-PMC scheme*, Technical report, arXiv:0805.2256, 2008, http://arxiv.org/abs/0805.2256.

[62] C. BIERNACKI, G. CELEUX, G. GOVAERT. *Exact and Monte Carlo Calculations of Integrated Likelihoods for the Latent Class Model*, Technical report, n^{o} 6609, Institut National de Recherche en Informatique et Automatique, 2008, http://hal.inria.fr/inria-00310137/fr/.

[63] G. CELEUX, A. GRIMAUD. *Approximation de systèmes complexes : linéarisation, approximation barycentrique, krigeage*, Technical report, Rapport de fin de contrat EDF, 2008.

[64] G. CELEUX, C. ROERO. *Étude de modèles discrets de durées de vie*, Technical report, Rapport de fin de contrat EDF, 2008.

[65] R. GENUER, J.-M. POGGI, C. TULEAU. *Random Forests: some methodological insights*, Technical report, n^{o} RR-6729, Institut National de Recherche en Informatique et Automatique, 2008, http://hal.inria.fr/inria-00340725/fr/.

[66] A. GRELAUD, C. ROBERT, J.-M. MARIN, F. RODOLPHE, J.-F. TALY. *ABC methods for model choice in Gibbs random fields*, Technical report, arXiv:0805.2256, 2008.

[67] F.-X. JOLLOIS, J.-M. POGGI, B. PORTIER. *Analyse statistique de la pollution par les particules en Haute-Normandie : étude descriptive*, Technical report, Rapport Air Normand (190 pages), 2008.

[68] F.-X. JOLLOIS, J.-M. POGGI, B. PORTIER. *Analyse statistique de la pollution par les particules en Haute-Normandie : modélisation et quantification des effets*, Technical report, Rapport Air Normand (161 pages), 2008.

[69] C. MAUGIS, G. CELEUX, M.-L. MARTIN-MAGNIETTE. *Variable selection in model-based clustering: a general variable role modeling*, Technical report, n^{o} RR-6744, Institut National de Recherche en Informatique et Automatique, 2008, http://hal.inria.fr/inria-00342108/fr/.

[70] C. MAUGIS, B. MICHEL. *A non asymptotic penalized criterion for Gaussian mixture model selection*, Technical report, n^{o} RR-6549, Institut National de Recherche en Informatique et Automatique, 2008, http://hal.inria.fr/docs/00/28/50/31/PDF/RR-6549.pdf.

[71] C. MAUGIS, B. MICHEL. *Slope heuristics for variable selection and clustering via Gaussian mixtures*, Technical report, n^{o} RR-6550, Institut National de Recherche en Informatique et Automatique, 2008, http://hal.inria.fr/docs/00/28/50/32/PDF/RR-6550.pdf.

[72] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Prévision par désagrégation : variantes, interprétation et effet de la taille des données*, Technical report, Rapport EDF (60 pages), 2008.

[73] N. VERZELEN. *High-dimensional Gaussian model selection on a Gaussian design*, Technical report, n^{o} RR-6616, Institut National de Recherche en Informatique et Automatique, 2008, http://hal.inria.fr/inria-00311412/fr/.