INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team select

Model Selection and Statistical Learning

Futurs

THEME COG

Table of contents

1. Team

Team Leader
Pascal Massart [Professor université Paris-Sud]

Team Vice-Leader
Gilles Celeux [DR Inria]

administrative assistant
Marie-Carol Lopes [TR, partially with the team TAO]

Staff member Inria
Jean-Michel Marin [CR, detached from Université Paris 9]

Staff member Université Paris-Sud
Christine Kéribin [Assistant Professor]
Marie-Anne Poursat [Assistant Professor, with the team since October 1, 2005]

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

Ph. D. student
Nicolas Bousquet [Inria grant]
Sohie Donnet [MESR grant]
Marc Lavarde [CIFRE grant]
Bertrand Michel [CIFRE grant]
Marie Sauvé [MESR grant]
Christine Tuleau [MESR grant]
Laurent Zwald [MESR grant]

Post-doctoral fellow
Guillaume Saint Pierre [Post Doc. fellowship until 31 October 2005]

Student intern
Romain François [March-Septemberr 2005]
Cathy Maugis [March-September 2005]

2. Overall Objectives

2.1. Overall Objectives

Our research domain is statistics. In the last decades, statistical methodology has received a lot of contributions. Many different methods and algorithms are available in current softwares of statistical learning. The user of these methods is facing the problem of choosing a relevant method for its data set and objective. The model selection problem is an important but difficult problem from both theoretical and practical point of views. Classical criteria of models selection, based on often unrealistic assumptions, are penalized minimum contrast criteria with fixed penalties. SELECT is aiming to provide efficient model selection criteria with data driven penalty terms. In this context, SELECT is expecting to improve the toolkit of statistical model selection criteria from both theoretical and practical aspects. Currently, SELECT is focusing its effort on variable selection in regression problems, 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 [25], [21], [15] .

3. Scientific Foundations

3.1. Model selection in Statistics

Keywords: *Abrupt changes*, *Bayesian inference*, *Concentration inequalities*, *Data-driven penalties*.

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

3.1.1. A non asymptotic view for model selection

An important purpose of SELECT is to build and analyze penalized log-likelihood model selection criteria efficient when the number of models in competition grows to infinity with the number of observations. Concentration inequalities are a key tool for that purpose and lead to propose data-driven penalty choices strategies. A major issue of SELECT consists of deepening the analysis of data-driven penalties both from the theoretical and the practical side. There is no universal way of calibrating penalties but there are several different rather general ideas that we want to develop, including heuristics derived from the Gaussian theory, special strategy for variable selection and making use of resampling methods [7], [1], [24], [37].

3.1.2. Taking account of the modelling purpose in model selection

Choosing a model is not only a difficult problem from the theoretical point of view. Model selection criteria have been conceived to answer the difficulty that the data probability distribution P is unknown. But, beyond technical difficulties which can occur when choosing a model, it can be fruitful to take into account the purpose of the model user to get reliable and useful models for statistical description or decision tasks. As noticed earlier, most of standard model selection criteria are assuming that P is belonging to one of the considered models without considering the modelling purpose. This point of view would be useful not only from the practical point of view, but also it could help to avoid or overcome theoretical difficulties. Moreover, taking into account the modelling purpose would produce flexible model selection criteria with data-driven penalties [9]. This point of view can be expected to be useful in supervised Classification and hidden structure models. Finally, it is worth to mention that an alternative Bayesian approach for taking the modelling purpose into account can be expected to be useful in that setting [32], [13].

3.1.3. 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 the fixed effect structure, covariance model for the random effects), models defined by Ordinary or Stochastic Differential Equations, left censured models, as well as design optimization for the trial itself [19], [22].

4. Application Domains

4.1. Application Domains

Keywords: *curves classification*, *phylogeny*, *reliability*.

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

4.1.1. Curves classification

An increasing interest is now evident in the field of classification and regression for complex data as curves, functions, spectra, time series and so on. Such questions naturally arise when each observation consists of values of explanatory variables which are not scalar valued but of functional nature. Classical questions widely

examined in Data Analysis are now revisited to take into account and to take advantage (if possible) of the functional nature of the data and to define original strategies [5]. Such questions are now related to a well identified domain called functional data analysis. Various applied problems strongly motivate this interest like longitudinal studies, analysis of fMRI data, spectral calibration, ...

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

4.1.2. Reliability

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

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

4.1.3. Phylogeny

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

5. Software

5.1. mixmod software

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

Participants: Gilles Celeux [correspondent], Romain François.

MIXMOD is developed with Christophe Biernacki, Florent Langrognet (Université de Franche-Comté) and Gérard Govaert (Université de Technologie de Compiègne). MIXMOD (MIXture MODelling) software fits mixture models to a given data set with either a clustering or a discriminant analysis purpose. A large variety of algorithms to estimate the mixture parameters are proposed (EM, Classification EM, Stochastic EM) and it is possible to combine them to lead to different strategies in order to get a sensible maximum of the likelihood (or completed likelihood) function. Moreover, different information criteria for choosing a parsimonious model (the number of mixture component, for instance), some of them favoring either a cluster analysis or a discriminant analysis view point, are included. Many Gaussian models for continuous variables and multinomial models for discrete variable are available. Written in C++, MIXMOD is interfaced with SCILAB and MATLAB. The software, the statistical documentation and also the user guide are available on

the Internet at the following address http://www-math.univ-fcomte.fr/mixmod/index.php. This year, Romain François has improved the MIXMOD CPU time performance by a factor of ten in implementing all the possible improvements previously detected by Florent Langrognet [6], [46].

5.2. monolix software

Keywords: *Non linear models, mixed effects.*

Participant: Marc Lavielle [correspondent].

MONOLIX is developed by the MONOLIX group chaired by Marc Lavielle and France Mentré (INSERM) (see 8.1.1) and implement the methodology described in 3.1.3. A first MATLAB version has already been produced and is available on the MONOLIX group website: http://www.math.u-psud.fr/~lavielle/monolix An R version of this software is in development.

6. New Results

6.1. Model selection in statistical learning

6.1.1. General tools for model selection

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

In collaboration with Lucien Birgé (Université Paris 6), Pascal Massart proposed a data driven calibration method using minimal and optimal penalties in the Gaussian frame work. This heuristic, which show encouraging results, is based on the fact that in simple cases, we have shown that the ratio between those two penalties is equal to two [7]. This way of designing data driven penalties are now compared to resampling strategies as bootstrap.

In collaboration with Carrenne Ludeña, Marc Lavielle has introduced a new method to estimate the number of significant coefficients in non ordered model selection problems. The method is based on a convenient random centering of the partial sums of the ordered observations. Based on $L-$statistics methods the consistency of the proposed estimator has been demonstrated. From the practical point of view, numerical experiments show the accuracy of the estimator [52].

Jean-Michel Marin has considered compatible prior distributions for model selection in a Bayesian setting. Bayesian model choice techniques are based on the model posterior distributions. This distribution is sensitive to the choice of the parameter prior distributions of each model. These prior distributions should then be compatible in some sense. Within the setting of comparing two nested models, the issue of compatible priors for Bayesian model determination has been developed. Jean-Michel Marin develop the idea that two priors are most compatible when the corresponding marginal distributions of the observations are closest to each other. For the comparison of two nested models, he proposes the following procedure: first, specify the prior distribution of the full model and, for the sub-model, choose the prior distribution minimizing the Kullback-Leibler divergence between the two marginal distributions. He gets some results for exponential families and linear regression models [36], [33].

6.1.2. Model selection in Classification

Participants: Gilles Celeux, Pascal Massart, Jean-Michel Poggi, Marie Sauvé, Christine Tuleau, Laurent Zwald.

Pascal Massart in collaboration with Stéphane Boucheron (Université Paris 7) is making use of concentration inequalities that they designed to validate the data driven calibration he proposed for model selection (see 6.1.1) [10].

Jean-Michel Poggi and Christine Tuleau have proposed a strategy for variables selection in classification from high dimensional data. This work is motivated by a real world problem: objectification which has been treated in a contract with Renault. It consists of explaining the subjective drivability using physical criteria

coming from signals measured during experiments. They suggest an approach for the discriminant variables selection taking advantage of the functional nature of the data. The problem is ill-posed since the number of explanatory variables is dramatically greater than the sample size. The strategy proceeds in three steps: a signal preprocessing, including wavelet denoising and synchronization, dimensionality reduction by compression using a common wavelet basis, and finally the selection of useful variables via a stepwise strategy involving successive applications of CART method [53], [54], [3], [44].

Marie Sauvé and Christine Tuleau have studied variable selection through CART method, both in the regression and binary classification frameworks. An exhaustive procedure on all the subsets is examined theoretically using the *à la Birgé-Massart* model selection viewpoint leading to "oracle" inequalities and allowing to perform variable selection by penalized empirical contrast. A simulation study and an application to real data illustrate the method. Further investigation seems to show that a suitably chosen partial examination of the huge collection of considered models leads to a tractable and efficient procedure [3], [43].

In collaboration with Guillaume Bouchard (Ranx Xerok), Gilles Celeux has pursued the analysis of their so-called Bayesian Entropy Criterion (BEC) to select a classification model taking into account the decisional purpose of a model by minimizing the integrated classification entropy. They precised its theoretical behavior in general situations and explored intensively its practical features with numerical applications in computer vision [9], [2].

6.1.3. Statistical learning methodology and theory

Participants: Gilles Celeux, Pascal Massart, Guillaume Saint Pierre, Laurent Zwald.

In collaboration with Gilles Blanchard (Berlin University), Laurent Zwald has developed and studied a new algorithm of classification, the Kernel Projection Machine (KPM). It relies on the fact that regularization can be obtained thanks to a dimensionality reduction method such as Kernel Principal Component Analysis (Kernel-PCA). The choice of the vector space involved in the KPM is guided by statistical studies of model selection using the penalized minimization of the empirical loss. The choice of the vector space involved in KPM is guided by statistical studies of model selection using penalized minimization of the empirical loss. This regularization procedure is closely connected with the finite dimensional projections studied in the statistical work of Birgé and Massart [7]. The utilization of KPM is more flexible than the one of SVM with comparable results. A statistical study of model selection has been provided, justifying partially the regularization step used in the KPM. They also exploit Kernel-PCA to get models fitting the structure of the input data [28], [4], [29], [45].

In collaboration with Magalie Fromont ((Université de Rennes) Chritine Tuleau has studied the k-Nearest Neighbor method for functional data. The k-Nearest Neighbor (kNN)) method is well documented in the finite dimensional case. For functional data, the procedure consists of applying standard kNN on the projections of the data in a suitable space of dimension d. The procedure then involves to select the dimension d and the number of neighbors k. The standard kNN and a slightly penalized version have been considered theoretically in this functional framework. In addition, some examples seem to show that the introduction of a small penalty stabilizes the selection process while preserving good performance [3] .

In collaboration with Christophe Biernacki (Université de Lille) and Gérard Govaert (UTC), Gilles Celeux has started to study a family of parsimonious latent class model to analyze multivariate discrete data sets. The estimation of the models parameters are considered through maximum likelihood and Bayesian inference. One of the interest of the discrete context is actually to allow for a comparison of both approaches since Bayesian inference is not embarrassed with technical difficulties [26].

Gilles Celeux and Guillaume Saint Pierre have developed a model-based method for classification from dissimilarities data sets. Their approach consists of assuming that the observed dissimilarity table is equal to Euclidean distance plus a measurement error. Unobserved attributes are therefore modelled using mixture of multivariate Gaussian distributions, each one of them corresponding to a different group. Allowing variance errors to be different between intra-group measurements and inter-group measurements, leads to a flexible classification method. Model parameters are estimated with the maximum likelihood method, and the allocation is done by assigning a new measure to the group maximizing the likelihood. This method can

be expected to be useful in classifying non Euclidean dissimilarity tables due to a measurement error. First numerical results show an encouraging behavior of the proposed algorithm, but further experiments are needed to delimit the application domain of this classification method [40] .

6.1.4. Statistical analysis of fMRI data

Participants: Sophie Donnet, Marc Lavielle.

A collaboration of SELECT with the SHFJ (Service Hospitalier Frederic Joliot, CEA) concerns the statistical analysis of fMRI time series. Indeed, an accurate estimation of the Hemodynamic Response Function (HRF) in functional Magnetic Resonance Imaging (fMRI) is crucial for a precise spatial and temporal estimate of the underlying neuronal processes. Biological literature suggests that response magnitude may vary with attention or ongoing activity. We therefore have tested a flexible model that allows for the variation of the magnitude of the HRF with time in a Bayesian framework. Under this model, the magnitude of the HRF evoked by a single event may vary across occurrences of the same type of event. This model is tested against a simpler model with a fixed magnitude using information theory. We develop a standard EM algorithm to identify the event magnitudes and the HRF. We test this hypothesis on a series of 32 regions (4 ROIS on eight subjects) of interest and find that the more flexible model is better than the usual model in most cases [16], [34], [48].

6.1.5. Adaptive importance sampling schemes

Participants: Gilles Celeux, Jean-Michel Marin.

6.1.5.1. Iterated importance sampling in missing data problems

Missing variable models are typical benchmarks for new computational techniques since the ill-posed nature of missing variable models offers a challenging ground for these techniques. This was the case for the EM algorithm and the Gibbs sampler, and this is also true for importance sampling schemes. Gilles Celeux and Jean-Michel Marin have developed a population Monte Carlo scheme taking advantage of the latent structure of the problem. The potential of this approach and its specifics in missing data problems are illustrated in settings of increasing difficulty, in comparison with existing approaches. The improvement brought by a general Rao–Blackwellisation technique has been analyzed [14], [47], [18] .

6.1.5.2. Asymptotic properties of a D-kernel Population Monte Carlo scheme

In the design of efficient simulation algorithms, one is often beset with a poor choice of proposal distributions. Although the performances of a given kernel can clarify how adequate it is for the problem at hand, a permanent on-line modification of kernels causes concerns about the validity of the resulting algorithm. While the issue is quite complex and most often intractable for MCMC algorithms, the equivalent version for importance sampling algorithms can be validated quite precisely. Jean-Michel Marin in collaboration with Randal Douc (École Polytechnique), Arnaud Guillin and Christian Robert (Université Paris 9) derive sufficient convergence conditions for a wide class of population Monte Carlo algorithms and show that Rao–Blackwellized versions asymptotically achieve an optimum in terms of a Kullback divergence criterion, while more rudimentary versions simply do not benefit from repeated updating [49], [50].

Moreover, variance reduction has always been a central issue in Monte Carlo experiments. Population Monte Carlo can be used to this effect, in that a mixture of importance functions, called a D-kernel, can be iteratively optimized to achieve the minimum asymptotic variance for a function of interest among all possible mixtures. The implementation of this iterative scheme is illustrated for the computation of the price of a European option in the Cox-Ingersoll-Ross model [50].

6.1.6. Reliability

Participants: Nicolas Bousquet, Gilles Celeux, Marc Lavarde, Pascal Massart.

In the framework of a contract with EDF concerning reliability, Nicolas Bousquet and Gilles Celeux have concentrated their effort on taking into account expert opinions in a relevant way for Bayesian analysis for highly censored failure and of small size data sets. They proposed objective criteria to measure the coherence and to calibrate prior information in regard of the actual feedback data. They also define elicitation strategies to give a fair importance to the prior information [30], [11], [27], [31].

In the framework of a contract with Altis and in collaboration of Patrick Pamphile (Orsay), Marc Lavarde and Pascal Massart have adapted and applied the penalized model selection criterion of Birgé-Massart (cf. 6.1.1) for an accelerated lifetime test problem.

6.1.7. Phylogeny

Participants: Christine Kéribin, Pascal Massart, Marie-Anne Poursat.

Pascal Massart and Marie-Anne Poursat developed a new mutation-selection approach for phylogenetics trees. The basic model is assuming a constant mutation rate and independent mutation events. This model is too rough to describe reality. Dependence is introduced with a fitness function extending mutation models for molecular evolution to mutation-selection models. The program consists of first introducing a semi-parametric graphical model involving the tree topology, branches lengths and mutation rates as unknown parameters and the fitness function as a non parametric unknown function. Those unknown quantities are estimated by maximizing a penalized log-likelihood. The possibility under investigation is to choose as penalty the squared norm of the fitness function in the self-reproducing Hilbert space associated to some chosen kernel K. For this choice of kernel, it is expected to take advantage of the intensive works developed in bioinformatics for learning with kernels from DNA sequences.

Christine Kéribin has developed the PMCov package which is dedicated to estimate the branch lengths and topological parameters of a covarion model, when the topology is fixed. Attention has been particularly taken in testing the validity of the program now available on Windows and Linux systems. A statistical test using simulations is under study in order to test a non covarion against a covarion model.

6.1.8. 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 have used Gaussian mixture models to cluster transcript profiling data from Arabidopsis functional genomics. The aim was to characterize differences between the AGP and LTP families [38]. This statistical analysis was a preliminary work and will be continued in the thesis that Cathy Maugis is starting on cluster analysis for transcript profiling data.

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

Participants: Nicolas Bousquet, Gilles Celeux, Marc Lavarde, Pascal Massart, Bertrand Michel, Jean-Michel Poggi, Marie Sauvé, Christine Tuleau.

SELECT has a contract with EDF regarding durability of nuclear components and aging mastership.

SELECT has a contract with Altis (Cifre grant) regarding accelerated lifetime tests in the production process of chips.

SELECT has a contract with IFP (Cifre grant of Bertrand Michel) on modelling exploitation process of a petrol basin. Purposes of this work are the classification of production profiles and developing model selection tools in the context of Poisson process [39].

The thesis of Christine Tuleau is supported by Renault and the thesis of Marie Sauvé is supported by Rhodia.

8. Other Grants and Activities

8.1. National Actions

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

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

8.1.1. MONOLIX Group

Participants: Sophie Donnet, Marc Lavielle.

The MONOLIX group (Modèles Non Linéaires à Effets Mixtes, http://www.math.u-psud.fr/~lavielle/monolix), chaired by Marc Lavielle and France Mentré (INSERM) is a multi-disciplinary group, that exchanges and develops activities in the field of mixed effect models. It involves scientists with varied backgrounds, interested both in the study and applications of these models: academic statisticians (theoretical developments), researchers from INSERM (applications in pharmacology) and INRA (applications in agronomy, animal genetics and microbiology), and scientists from the medical faculty of Lyon-Sud University (applications in oncology). These collaborations have already translated in two co-chaired PhDs (INSERM/University and INRA/University), several papers, the organization of special sessions in two congresses (SMAI 2005, SFdS 2005) and participation in several international conferences (Lyon 2004, INRA 2005, PAGE 2004, PAGE 2005).

8.1.2. Action incitative DataHighDim

Participant: Gilles Celeux.

This ACI started in September 2003. Partners of ACI DataHighDim are laboratory CLIPS of UJF and laboratory LIS, INPG in Grenoble, SELECT team of INRIA, laboratory DICE, UCL in Louvain la Neuve and laboratory LDG, CEA Bruyères le Châtel. DataHighDim is concerned with exploratory and decisional analysis in high dimensions. This year three meetings of the group has been organized. The post-doc position of Guillaume Saint Pierre took place in this framework.

8.2. European actions

Gilles Celeux and Pascal Massart are members of the PASCAL (Pattern Analysis, Statistical Learning and Computational Learning) network. In this framework, Gilles Celeux and Jean-Michel Marin spent a week in Glasgow in the statistical department of Glasgow University. Both of them gave a conference during their stay at Glasgow in September 2005.

9. Dissemination

9.1. Scientific Communauty animation

Pascal Massart is associated editor of *Annales de l'IHP*, *Journal of the European Mathematical Society*, *Journal de la SFDS* and *ESAIM Proceedings*.

Gilles Celeux is associated editor of *Statistics and Computing*.

Pascal Massart was invited at the Probility meeting on Banach spaces in Vilnius in June 2005.

Gilles Celeux was invited speaker in the annual working group on Model-based cluster analysis of the University of Washington in Seattle in July 2005 and in the meeting "Data mining and Statistical Learning" of Niot in May 2005.

Pascal Massart is member of the scientific council of Euradom and of the working group on "le rôle des mathématiques dans le monde contemporain" of the french *Académie des Sciences*.

Gilles Celeux ans Pascal Massart have been members of the Progam committee of the 37^{th} Journées of statistiques of Pau.

9.2. Enseignement

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

10. Bibliography

Books and Monographs

[1] P. MASSART. *Concentration inequalities and model selection*, to appear, Lecture Notes in Mathematics, Springer, 2005.

Doctoral dissertations and Habilitation theses

[2] G. BOUCHARD. *Les modèles génératifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle*, Ph. D. Thesis, Université Joseph Fourier, Grenoble, May 2005.

[3] C. TULEAU. *Sélection de variables pour la discrimination en grande dimension et classification de données fonctionnelles*, Ph. D. Thesis, Université Paris-Sud, December 2005.

[4] L. ZWALD. *Performances statistiques d'algorithmes d'apprentissage : Kernel Projection Machine et Analyse en Composantes Principales à noyau*, Ph. D. Thesis, Université Paris-Sud, November 2005.

Articles in refereed journals and book chapters

[5] M. AMINGHAFARI, N. CHÈZE, J.-M. POGGI. *Multivariate denoising using wavelets and principal components*, in "Computational Statistics and Data Analysis", to appear, 2005.

[6] C. BIERNACKI, G. CELEUX, G. GOVAERT, F. LANGROGNET. *Model-based cluster analysis and discriminant analysis with the MIXMOD software*, in "Computational Statistics and Data Analysis", to appear, 2005.

[7] L. BIRGÉ, P. MASSART. *Minimal penalties for Gaussian model selection*, in "Probability thory and Related Fields", to appear, 2005.

[8] R. BISCAY, M. LAVIELLE, C. LUDEÑA. *Estimation of nonparametric autoregressive time series models under dynamical constraints*, in "Jour. of Time Series Anal.", vol. 26, n^o 3, 2005, p. 371–397.

[9] G. BOUCHARD, G. CELEUX. *Selection of generative models in Classification*, in "IEEE Trans. on PAMI", to appear, 2005.

[10] S. BOUCHERON, O. BOUSQUET, G. LUGOSI, P. MASSART. *Moment inequalities for functions of independent random variables*, in "Annals of Probability", vol. 33, 2005, p. 514-560.

[11] N. BOUSQUET. *An introduction to a new way of choosing and aggregating prior distributions for Weibull inference in an industrial durability/reliability context*, Special Issue, (to appear), 2005.

[12] G. CELEUX, F. CORSET, A. LANNOY, B. RICARD. *Designing a Bayesian network03 for preventive maintenance from expert opinion in a rapid and reliable way*, in "Reliability Engineering & System Safety", to appear, 2005.

[13] G. CELEUX, F. FORBES, C. P. ROBERT, D. M. TITTERINGTON. *Deviance information criteria for missing data models*, in "Bayesian Analysis", to appear, 2005.

[14] G. CELEUX, J.-M. MARIN, C. ROBERT. *Iterated importance sampling in missing data problems*, in "Computational Statistics and Data Analysis", to appear, 2005.

[15] G. CELEUX, O. MARTIN, C. LAVERGNE. *Mixture of linear mixed models - Application to repeated data clustering*, in "Statistical Modelling", vol. 5, 2005, p. 243-267.

[16] S. DONNET, M. LAVIELLE, J. POLINE. *Are fMRI event related response constant in time?*, in "NeuroImage (to appear)", 2005.

[17] S. GEY, J.-M. POGGI. *Boosting and Instability for regression trees*, in "Computational Statistics and Data Analysis", to appear, 2005.

[18] A. GUILLIN, J.-M. MARIN, C. ROBERT. *Estimation bayésienne approximative par échantillonnage préférentiel*, in "Revue de Statistique Appliquée", vol. LIII, n° 1, 2005, p. 79-85.

[19] E. KUHN, M. LAVIELLE. *Maximum likelihood estimation in nonlinear mixed effects models*, in "Computational Statistics and Data Analysis", vol. 49, n° 4, 2005, p. 1020–1038.

[20] M. LAVIELLE, C. LÉVY-LEDUC. *Semi parametric estimation of the frequency of unknown periodic functions*, in "IEEE Trans. on Signal Processing", vol. 53, n° 7, 2005, p. 2306–2314.

[21] M. LAVIELLE. *Using penalized contrasts for the change-points problem*, in "Signal Processing", vol. 85, n° 8, 2005, p. 1501–1510.

[22] D. MAKOWSKI, M. LAVIELLE. *Using SAEM to estimate parameters of models of response to applied fertilizer*, in "Joural of Agricultural, Biological, and Environmental Statistics (to appear)", 2005.

[23] J.-M. MARIN, K. MENGERSEN, C. ROBERT. *Bayesian modelling and inference on mixtures of distributions*, in "Handbook of Statistics", 25, Elsevier-Sciences, 2005.

[24] P. MASSART, E. NEDELEC. *Risk bounds for statistical learning*, in "Annals of Statistics", to appear, 2005.

[25] F. PICARD, S. ROBIN, M. LAVIELLE, C. VAISSE, J. DAUDIN. *A statistical approach for CGH microarray data analysis*, in "BMC Bioinformatics", vol. 6, n° 27, 2005.

Publications in Conferences and Workshops

[26] C. BIERNACKI, G. CELEUX, G. GOVAERT. *Parsimonious Latent Class Models*, in "3rd IASC world conference on Computational Statistics & Data Analysis, Limassol, Cyprus", October 2005.

[27] F. BILLY, N. BOUSQUET, G. CELEUX. *Modelling and eliciting expert knowledge with fictitious data*, in "Proceedings of the Workshop on the use of Expert Judgement for decision-making, CEA Cadarache", 2005.

[28] G. BLANCHARD, P. MASSART, R. VERT, L. ZWALD. *Kernel projection machine: a new tool for Pattern Recognition*, in "Adavance in Neural Information Processing Systems 17", MIT Press, Cambridge, 2005, p. 1649-1656.

[29] G. BLANCHARD, L. ZWALD. *On the Convergence of Eigenspaces in Kernel Principal Component Analysis*, in "Proceedings of the 19th. Neural Information Processing System (NIPS 2005)", MIT Press, 2005.

[30] N. BOUSQUET. *Choosing prior distributions for Weibull inference in a durability context: some propositions*, in "Proceedings International Symposium on Stochastic Models in Reliability, Safety, Security and Logistics, Beer Sheva, Israel", February 2005.

[31] N. BOUSQUET, G. CELEUX, E. REMY. *A protocol for integrating FED and expert data in a study of durability*, in "Proceedings of the Workshop on the use of Expert Judgement for decision-making, CEA Cadarache", 2005.

[32] G. CELEUX. *Sélection de modèles latents à information manquante minimale*, in "Journées modèles à données manquantes, Marne-la-Vallée", January 2005.

[33] G. CELEUX, J.-M. MARIN. *Sélection bayesienne de variables en régression linéaire via des lois a priori compatibles*, in "Data Mining et Apprentissage Statistique, Niort", Mai 2005.

[34] S. DONNET. *Inversion de données IRMf. Estimation et sélection de modèles*, in "37èmes Journées de Statistique organisées par la Société Française de Statistique, Pau", June 2005.

[35] J.-M. MARIN. *Bayesian Modeling and Inference on Mixtures of Distributions*, in "Journées modèles à données manquantes, Marne-la-Vallée", January 2005.

[36] J.-M. MARIN. *Sélection bayésienne de variables en régression linéaire via des lois a priori compatibles*, in "XXXVII èmes Journées de Statistique, Pau", June 2005.

[37] P. MASSART. *A non asymptotic theory for model selection*, in "4th European Congress of Mathematicians", A. LAPTEV (editor). , 2005, p. 309-323.

[38] C. MAUGIS, S. AUBOURG, J.-P. RENOU, G. CELEUX, M.-L. MARTIN-MAGNIETTE. *Utilisation des modèles de mélange de distribution de probabilités pour la classification de données transcriptome sur les familles AGP et LTP*, in "Journées du Réseau Français des Parois, Rouen", september 2005.

[39] B. MICHEL. *Hypothèses retenues pour la modélisation de la production d'hydrocarbures au sein d'un bassin pétrolier*, October 2005.

[40] G. SAINT-PIERRE, G. CELEUX. *Model based classification with dissimilarities : A maximum likelihood approach*, in "Panorama des recherches incitatives en sciences et technologies de l'information et de la communication, Bordeaux", November 2005.

[41] G. SAINT-PIERRE. *Algorithme MCMC à sauts réversibles pour les mélanges gaussiens multivariés*, in "37èmes Journées de Statistique organisées par la Société Française de Statistique, Pau", June 2005.

[42] G. SAINT-PIERRE. *Méthodes MCMC et mélanges gaussiens*, in "Premières rencontres de jeunes statisticiens, Aussois", August 2005.

[43] M. SAUVÉ, C. TULEAU. *Sélection de variables avec CART*, in "XXXVII èmes Journées de Statistique, Pau", June 2005.

[44] C. TULEAU. *Classification en grande dimension - Sélection de variables*, in "Premières rencontres de jeunes statisticiens, Aussois", August 2005.

[45] L. ZWALD. *Statistical Performances of some learning algorithm : Kernel projection Machine and Kernel Principal Component Analysis*, December 2005.

Internal Reports

[46] C. BIERNACKI, G. CELEUX, G. GOVAERT, F. LANGROGNET. *Model-based Cluster and Discriminant Analysis with the MIXMOD software*, Technical report, nº RT-0302, Institut National de Recherche en Informatique et Automatique, 2005, http://www.inria.fr/rrrt/rt-0302.html.

[47] G. CELEUX, J.-M. MARIN, C. ROBERT. *Iterated importance sampling in missing data problems*, Technical report, nº RR-5534, Institut National de Recherche en Informatique et Automatique, 2005, http://www.inria.fr/rrrt/rr-5534.html.

[48] S. DONNET, A. SAMSON. *Estimation of parameters in missing data models defined by differential equations*, Technical report, nº 2005-29, Laboratoire de mathématiques. Université Paris XI. Orsay, 2005.

[49] R. DOUC, A. GUILLIN, J.-M. MARIN, C. ROBERT. *Convergence of adaptative sampling schemes*, Technical report, nº RR-5485, Institut National de Recherche en Informatique et Automatique, 2005, http://www.inria.fr/rrrt/rr-5485.html.

[50] R. DOUC, A. GUILLIN, J.-M. MARIN, C. ROBERT. *Minimum variance importance sampling via Population Monte Carlo*, Technical report, nº RR-5699, Institut National de Recherche en Informatique et Automatique, 2005, http://www.inria.fr/rrrt/rr-5699.html.

[51] W. KENDALL, J.-M. MARIN, C. ROBERT. *Brownian confidence bands on Monte Carlo output*, Technical report, nº RR-5436, Institut National de Recherche en Informatique et Automatique, 2004, http://www.inria.fr/rrrt/rr-5436.html.

[52] M. LAVIELLE, C. LUDEÑA. *Random thresholds for linear model selection*, Technical report, nº RR-5572, Institut National de Recherche en Informatique et Automatique, 2005, http://www.inria.fr/rrrt/rr-5572.html.

[53] J.-M. POGGI, C. TULEAU. *Classification supervisée en grande dimension: application à l'agrément de conduite*, Technical report, nº 28, Prépublication Université Orsay, 2005.

Miscellaneous

[54] J.-M. POGGI, C. TULEAU. *Méthodologie de hiérarchisation de mesures et d'identification de plages pertinentes pour objectiver une prestation. Application au décollage BVR.*, 2005.