INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team select

Model Selection and Statistical Learning

Futurs

THEME COG

Activity Report

2004

Table of contents

1. Team

Head of project-team

Pascal Massart [Professeur université Paris-Sud]

Vice-head of project team

Gilles Celeux [DR Inria]

administrative assistant

Marie-Carol Lopes [TR, à temps partiel dans l'équipe]

Staff member Inria

Jean-Michel Marin [CR, détaché de l'université Paris 9 depuis le 01/09/04]

Staff member CNRS

Jean-Michel Loubes [CR]

Staff member Université Paris-Sud

Christine Kéribin [Maître de conférences]

Staff member Université Pari 5

Marc Lavielle [Professeur]
Jean-Michel Poggi [Professeur]

Ph. D. student

Guillaume Bouchard [allocataire Inria, en partie dans le projet Lear]
Olivier Bousquet [allocataire Inria]
Marc Lavarde [allocataire CIFRE]
Marie Sauvé [allocataire MESR]
Christine Tuleau [allocataire MESR]
Laurent Zwald [allocataire MESR]

Post-doctoral fellow

Guillaume Saint Pierre [Post Doctorant Inria depuis le 1 novembre 2004]

Student intern

Romain Goutaland [mars-août 2004]

2. 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, abrupt changes detection, hidden structure models and supervised classification. Its domain of application concern reliability, curves classification, phylogeny analysis and classification in genetics.

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 [21], [20].

3.1.2. Multiple Change points detection

The change-point problem is important in many applications, and has been well-studied for more than forty years. We are focusing on the *a posteriori* problem which consists of recovering the configuration of change-points using the whole observed series. This problem has many potential applications. It can have a central role in the analysis of seismic signals, ECG or EEG signals for instance. We are developing a procedure is aiming to detect all the change points simultaneously by minimizing a *penalized contrast*. Different contrast functions can be used according to the problem at hand. But, the difficult problem which remains partly open is to define and analyze proper data driven strategies for calibrating the penalty. Some calibration strategies have already been proposed, But there is still some hard work to be done in order to validate those methods.

3.1.3. 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 [16]. 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.

4. 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 conventions 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 [1], [2]. 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. Of course, in addition to classical questions like the choice of the number of clusters, the choice of the norm or pseudo-norm to measure the distance between two observations, the choice of an observation or a point to reduce a cluster to the most representative observation and so on, 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 convention 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 [17].

The other theme of research in which SELECT is involved concern changes in a reliability process. It comes from a convention 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 aluminium 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.

The model that we consider is 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). We are working to elucidate identifiability conditions and to analyze those conditions from the practical viewpoint. In this research, we are also interested to compare non nested models [19].

5. Software

5.1. mixmod software

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

Participant: Gilles Celeux [correspondant].

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. can be considered according to different assumptions on the component variance matrix eigenvalue decomposition. 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.

6. New Results

6.1. Model selection in statistical learning

6.1.1. Concentration inequalities

Participant: Pascal Massart.

Joint work with Stéphane Boucheron (LRI, Orsay), Olivier Bousquet (Max Planck Institute, Tuebingen) and Gabor Lugosi (Pompeu Fabra, Barcelona). New inequalities for functions of independent random variables have been designed. They prove to be a versatile tool in a wide range of applications. In particular, the Talagrand's exponential inequality for Rademacher chaos of order two has been generalized to any order. Applications for other complex functions of independent random variables, such as suprema of Boolean polynomials which include, as special cases, subgraph counting problems in random graphs have been considered [4].

6.1.2. Model selection in Classification

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

Guillaume Bouchard and Gilles Celeux have proposed a new criterion, the 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. It provides an interesting alternative to the cross validated error rate which is highly time consuming. The asymptotic behavior of BEC criterion has been studied. Numerical experiments on both simulated and real data sets show that BEC is performing better than the classical BIC criterion to select a model minimizing the classification error rate [26].

In collaboration with Lucien Birgé (Paris 6), Pascal Massart has analyzed in a precise way what kind of penalties should be used in order to perform model selection via the minimization of a penalized least squares criterion within some general Gaussian framework [24].

Marie Sauvé and Christine Tuleau [22] are studying a variable selection procedure based on CART in the Gaussian regression framework and in the classification framework. This CART (Classification And Regression Trees) algorithm is a popular algorithm which builds a piecewise constant estimator of a regression function or a classifier from a training sample of observations [7].

Jean-Michel Poggi and Christine Tuleau are designing classification rules using CART, wavelet-based compression and denoising for measuring the comfort of driving. It a an applied study supported by Renault.

6.1.3. Statistical learning methodology and theory

Participants: Guillaume Bouchard, Pascal Massart, Laurent Zwald.

In collaboration with Gilles Blanchard and Régis Vert (Orsay) Pascal Massart and Laurent Zwald have introduced a new kernel algorithm for pattern recognition. They started from a study of the regularization properties of Kernel Principal Component Analysis (KPCA) within the classification framework. KPCA has been previously used as a pre-processing step of support vector machine (SVM) but this method is somewhat redundant from a regularization point of view and they propose a new algorithm called *Kernel Projection Machine* to avoid this redundancy, based on an analogy with the statistical framework of regression for a Gaussian white noise model. Preliminary experimental results show that this algorithm reaches the same performances as SVM [15].

Moreover in collaboration of Gilles Blanchard and Olivier Bousquet, Laurent Zwald has studied the properties of the eigenvalues of Gram matrices in a non-asymptotic setting. Using local Rademacher averages, they provide data-dependent tight bounds for their convergence toward eigenvalues of the corresponding kernel operator. They perform these computations in a functional analytic framework which allows to deal implicitly with reproducing kernel Hilbert spaces of infinite dimension. This can have applications to various kernel algorithms. Focusing on KPCA they get sharp excess risk bounds for the reconstruction error [14].

In collaboration with Bill Triggs (team LEAR, Inria), Guillaume Bouchard has proposed and studied a method providing a compromise between a generative classifier (modelling the joint distribution of the groups and the descriptive variables) and a discriminative classifier (modelling the conditional distribution of the groups knowing the descriptive variables) [18].

6.1.4. Detection of abrupt changes

Participants: Marc Lavielle, Romain Goutaland.

A methodology for model selection based on a penalized contrast has been developed by Marc Lavielle. This methodology is applied to the change-point problem, for estimating the number of change points and their location. An adaptive choice of the penalty function has been proposed for automatically estimating the number of change points. In a Bayesian framework, the posterior distribution of the change-point sequence as a function of the penalized contrast has been defined. Monte Carlo Markov chains (MCMC) procedures are available for sampling this posterior distribution. The parameters of this distribution are estimated with a stochastic version of EM algorithm (SAEM) [27].

Moreover, in collaboration with researchers of INA, this methodology has been used for analyzing Micro-array-CGH experiments which aim at detecting and mapping chromosomal imbalances, by hybridizing targets of genomic DNA from a test and a reference sample [28].

On an other hand, a preliminary effort with the training course of Romain Goutaland has been done to transform the **DCPC** (Detection of Changes using Penalized Contrasts) procedure [27] in a general software for multiple change points detection. Different contrast functions has been considered in the model of software coded in C++ by Romain Goutaland.

6.1.5. Reliability

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

In collaboration with Henri Bertholon (CNAM, Paris), Nicolas Bousquet and Gilles Celeux have proposed a simple competing risk distribution as a possible alternative to the Weibull distribution in lifetime analysis. This distribution is the minimum between exponential and Weibull distributions. The motivation was to take account of both accidental and aging failures in lifetime data analysis. The estimation of the parameters of this distribution are considered through maximum likelihood and Bayesian inference. Decision tests to choose between an exponential, Weibull and this competing risk distribution have been proposed [23] [13], [12].

In the framework of a convention with EDF, they have investigated the ability of adaptive importance sampling schemes to lead to efficient estimators of posterior reliability distributions from highly censored data. And, they have proposed several strategies for eliciting expert opinions to deal with informative Bayesian inference in a proper way for competing risk models involving Weibull distributions.

In the framework of a convention 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. 3.1) for an accelerated lifetime test problem.

6.1.6. Phylogeny

Participant: Christine Kéribin.

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. A statistical test using simulations will be soon proposed in order to test a non covarion against a covarion model.

7. Contracts and Grants with Industry

7.1. EDF

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

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

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

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. Actions nationales

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

8.1.1. Action incitative MIST-R

Participants: Gilles Celeux, Jean-Michel Loubes.

This ACI started in September 2004. Partners of ACI MIST-R are CNRS (section 01), INRIA (SELECT and TAO teams), Paris-Sud University (LRI and mathematical department), University Paul Sabatier of Toulouse (laboratory of statistics et probability) and INRETS (laboratory GRETIA). The coordinator is Jean-Michel Loubes.

MIST-R is concerned with cars traffic prevision. The statistical methods SELECT planned for this purpose are curves classification, mixture analysis of semi parametric distributions, distance tables analysis and variables selection procedures.

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 first meeting in Grenoble has been entirely devoted to the presentation by Gilles Celeux of probabilistic models for the classification of distances tables and the Bayesian estimation of those models.

8.2. Actions européennes

Gilles Celeux and Pascal Massart are participants of the PASCAL (Pattern Analysis, Statistical Learning and Computational Learning) network.

9. Dissemination

9.1. Animation de la Communauté scientifique

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 has been plenary invited speaker of *IFCS2004* in Chicago. Pascal Massart has been plenary invited speaker of the meeting *Mathematical foundations of statistical learning* in Barcelona and invited to the

4th *ECM* meeting in Stockholm. He gave a series of five conferences on model selection in Hilversum (The Netherlands).

Guillaume Bouchard received *Gold* $\lambda\mu$ of best Ph. D. Student thesis for his work in [17].

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] M. MISITI, Y. MISITI, G. OPPENHEIM, J.-M. POGGI. *Matlab WAvelet Toolbox (Version 3.0): Tutorial and Reference Guide*, The Mathworks, Natick, USA, 2004.

Articles in referred journals and book chapters

[2] M. AMINGHAFARI, N. CHÈZE, J.-M. POGGI. *Multivariate denoising using wavelets and principal components*, in "Computational Statistics and Data Analysis", to appear, 2004.

[3] R. BISCAY, M. LAVIELLE, C. LUDENA. *Estimation of nonparametric autoregressive time series models under dynamical constraints*, in "IEEE Trans. on Signal Processing", to appear, 2004.

[4] S. BOUCHERON, O. BOUSQUET, G. LUGOSI, P. MASSART. *Moment inequalities for functions of independent random variables*, in "Annals of Probability", to appear, 2004.

[5] G. CELEUX, J. NASCIMENTO, J. MARQUES. *Learning switching dynamic models for objects tracking*, in "Pattern Recognition", vol. 37, 2004, p. 1841-1853.

[6] J.-B. DURAND, L. BOZZI, G. CELEUX, C. DERQUENNE. *Analyse de courbes de consommation électrique par chaines de Markov cachées*, in "Revue de Statistique Appliquée", vol. 52, 2004, p. 71-91.

[7] S. GEY, J.-M. POGGI. *Boosting and Instability for regression trees*, in "Computational Statistics and Data Analysis", to appear, 2004.

[8] E. KUHN, M. LAVIELLE. *Coupling a stochastic approximation version of EM with a MCMC procedure*, in "Journal of Times Series Analysis", to appear, 2004.

[9] E. KUHN, M. LAVIELLE. *Maximum likelihood estimation in nonlinear mixed effects models*, in "Computational Statistics and Data Analysis", to appear, 2004.

[10] M. LAVIELLE, C. LÉVY-LEDUC. *Semiparametric estimation of the frequency of unknown periodic functions. Application to laser vibrometry signals*, in "IEEE Trans. on Signal Processing", to appear, 2004.

[11] J.-M. LOUBES, P. MASSART. *Discussion to Least Angle Regression*, in "Annals of Statistics", vol. 32, 2004, p. 476-482.

Publications in Conferences and Workshops

[12] H. BERTHOLON, N. BOUSQUET, G. CELEUX. *A competing risk lifetime model*, in "Mathematical Methods in Reliability, Santa Fe", June 2004.

[13] H. BERTHOLON, N. BOUSQUET, G. CELEUX. *Un modèle de durée de vie à risques concurrents*, in "36èmes Journées de Statistique organisées par la Société Française de Statistique, Montpellier", June 2004.

[14] G. BLANCHARD, O. BOUSQUET, L. ZWALD. *Statistical Properties of Kernel Principal Component Analysis*, in "COLT 2004", August 2004.

[15] G. BLANCHARD, P. MASSART, R. VERT, L. ZWALD. *Kernel Projection Machine: a New Tool for Pattern Recognition*, in "NIPS 2004", December 2004.

[16] G. BOUCHARD, G. CELEUX. *Model Selection in Classification*, in "IFCS2004, Chicago", July 2004.

[17] G. BOUCHARD, G. CELEUX, F. BILLY, F. JOSSE. *Réactualisation bayésienne dún modèle de dégradation en fonction du retour d'expérience*, in "Proceedings of $\lambda\mu$14, Bourges", vol. 1, October 2004, p. 39-46.

[18] G. BOUCHARD, B. TRIGGS. *The Tradeoff between Generative and Discriminative classifiers*, in "COMPSTAT2004, Prague", August 2004, p. 721-728.

[19] C. KÉRIBIN. *Test de modèle en phylogénie*, in "Journées MAS, Nancy", September 2004.

[20] P. MASSART. *A non asymptotic theory for model selection*, in "Proceedings of Mathematical Foundations of Stattistical Learning, Stockholm", July 2004.

[21] P. MASSART. *Optimal Selection in Classification*, in "Proceedings of Mathematical Foundations of Statistical Learning, Barcelona", June 2004.

[22] J.-M. POGGI, C. TULEAU. *Classification supervisée en grande dimension: application à l'agrément de conduite*, in "36èmes Journées de Statistique organisées par la Société Française de Statistique, Montpellier", June 2004.

Internal Reports

[23] H. BERTHOLON, N. BOUSQUET, G. CELEUX. *An alternative competing risk model to the Weibull distribution in lifetime data analysis*, Technical report, n° RR-5265, Institut National de Recherche en Informatique et Automatique, 2004, http://www.inria.fr/rrrt/rr-5265.html.

[24] L. BIRGÉ, P. MASSART. *Minimal penalties for Gaussian model selection*, Preprint, Technical report, 2004.

[25] G. BLANCHARD, O. BOUSQUET, P. MASSART. *Statistical performance of support vector machine*, Preprint, Technical report, 2004.

[26] G. BOUCHARD, G. CELEUX. *Model selection in supervised classification*, Technical report, n^o RR-5391, Institut National de Recherche en Informatique et Automatique, 2004, http://www.inria.fr/rrrt/rr-5391.html.

[27] M. LAVIELLE. *Using penalized contrasts for the change-point problem*, Technical report, n^o RR-5339, Institut National de Recherche en Informatique et Automatique, 2004, http://www.inria.fr/rrrt/rr-5339.html.

[28] F. PICARD, S. ROBIN, M. LAVIELLE, C. VAISSE, J.-J. DAUDIN. *A statistical approach for CGH microarray data analysis*, Technical report, n^o RR-5139, Institut National de Recherche en Informatique et Automatique, 2004, http://www.inria.fr/rrrt/rr-5139.html.