IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2011

Project-Team SELECT

Model selection in statistical learning

IN COLLABORATION WITH: Laboratoire de mathématiques d'Orsay de l'Université de Paris-Sud (LMO)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Optimization, Learning and Statistical Methods

Table of contents

Project-Team SELECT

Keywords: Data Analysis, Data, Machine Learning, Statistical Learning, Decision Methods

1. Members

Research Scientists

Gilles Celeux [Team Vice-Leader, Senior Researcher INRIA, HdR]
Erwan Le Pennec [Junior Researcher INRIA]

Faculty Members

Pascal Massart [Team Leader, Professor Université Paris-Sud, HdR]
Christine Keribin [Associate Professor]
Jean-Michel Poggi [Professor Université Paris 5, HdR]

External Collaborator

Yves Auffray [Dassault]

PhD Students

Vincent Brault [MESR grant]
Mohammed El Anbari [France-Marocco grant]
Rémi Fouchereau [MESR grant]
Shuai Fu [EDF-INRIA Cifre grant]
Robin Genuer [MESR grant]
Caroline Meynet [MESR grant]
Nelo Molter Magalães [MESR grant]
Lucie Montuelle [MESR grant]
Clément Levrard [MESR grant]

Post-Doctoral Fellows

Caroline Bérard [ATER]
Jairo Cugliari-Duhalde [Post-Doc]

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, 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) and population genetics.

3. Scientific Foundations

3.1. General presentation

We learned from the applications we treated that some assumptions which are currently used in asymptotic theory for model selection are often irrelevant in practice. For instance, it is not realistic to assume that the target belongs to the family of models in competition. Moreover, in many situations, it is useful to make the size of the model 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 modeling 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. 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.

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, although we also have several more academic collaborations, e.g. genomics, genetics and neuroimaging.

4.2. Curves classification

The field of classification for complex data as curves, functions, spectra and time series is important. Standard data analysis questions are being revisited to define new strategies that take the functional nature of the data into account. Functional data analysis addresses a variety of applied problems, including longitudinal studies, analysis of fMRI data and spectral calibration.

We are focusing on unsupervised classification. In addition to standard questions 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 needs to be design efficient anytime algorithms.

4.3. Computer Experiments and Reliability

Since several years, SELECT has collaborations with EDF-DER *Maintenance des Risques Industriels* group. An important theme concerns the resolution of inverse problems using simulation tools to analyze incertainty in highly complex physical systems. A collaboration on an analogous topic is developed with Dassault Aviation.

The other major theme concerns probabilistic modeling in fatigue analysis in the context of a research collaboration with SAFRAN an high-technology group (Aerospace propulsion, Aicraft equipment, Defense Security, Communications).

4.4. 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 have been co-supervised by SELECT and Neurospin researchers (Merlin Keller 2006-2009 and Vincent Michel 2007-2010). 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.

4.5. Analysis of genomic data

For the past few years SELECT has collaborated with Marie-Laure Martin-Magniette (URGV) for the analysis of genomic data. An important theme of this collaboration is using statistically sound model-based clustering methods to discover groups of co-expressed genes from microarray and high-throughput sequencing data. In particular, identifying biological entities that share similar profiles across several treatment conditions, such as co-expressed genes, may help identify groups of genes that are involved in the same biological processes.

4.6. Environment

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 and forecasted PM10 pollution in Rouen area on six different monitoring sites to quantify the effects of variables of different types, mainly meteorological versus other pollutant measurements. Some 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.

5. Software

5.1. MIXMOD software

Participants: Gilles Celeux [Correspondant], Erwan Le Pennec.

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.mixmod.org.

Since this year, MIXMOD has a proper graphical user interface (Version 1) which has been presented at the MIXMOD day in Lyon in December 2010.A version of MIXMOD in R is forthcoming.

Erwan Le Pennec with the help of Serge Cohen has proposed a spatial extension in which the mixture weights can vary spatialy.

6. New Results

6.1. Model selection in Regression and Classification

Participants: Gilles Celeux, Mohammed El Anbari, Clément Levrard, Robin Genuer, Erwan Le Pennec, Lucie Montuelle, Pascal Massart, Caroline Meynet, Jean-Michel Poggi.

Erwan Le Pennec continues his work with Serge Cohen (IPANEMA Soleil) on hyperspectral image segmentation based on a spatialized Gaussian Mixture Model. They derive, and implement within MIXMOD, an efficient minimization algorithm combining EM algorithm, dynamic programming and model selection[37]. They have applied this technique to analyze ancient material[9] This scheme is supported by a theoretical work on conditional density estimation[40]. In the framework of her PhD, Lucie Montuelle has studied some extension to this model to spatiay varying logistic weights.

In collaboration with Marie-Laure Martin-Magniette (URGV et UMR AgroParisTech/INRA MIA 518) and Cathy Maugis (INSA Toulouse) has extended their variable selection procedure for model-based clustering and supervised classification to deal with high dimensional data sets with a backward selection procedure which is more efficient that the previous forward selection procedure in this context. [17]. Moreover they have shown the advantage of the model-based approach over a geometrical approach to select variable for clustering [13]. These variable selection procedures are in particular used for genomics applications which is the result of a collaboration with researchers of of URGV (Evry Genopole).

Caroline Meynet provided an ℓ_1-oracle inequality satisfied by the Lasso estimator with the Kullback-Leibler loss in the framework of a finite mixture of Gaussian regressions model for high-dimensional heterogeneous data where the number of covariates may be much larger than the sample size. In particular, she has given a condition on the regularization parameter of the Lasso to obtain such an oracle inequality. This oracle inequality extends the ℓ_1-oracle inequality established by Massart and Meynet [16] in the homogeneous Gaussian linear regression case. It is deduced from a finite mixture Gaussian regression model selection theorem for ℓ_1-penalized maximum likelihood conditional density estimation, which is inspired from Vapnik's method of structural risk minimization and from the theory on model selection for maximum likelihood estimators developed by Massart.

From an practical point of view, Caroline Meynet has introduced a procedure to select variables in model-based clustering in a high-dimensional context. In order to tackle with the problem of high-dimension, she has proposed to first use the Lasso in order to select different sets of variables and then estimate the density by a standard EM algorithm by reducing the inference to the linear space of the selected variables by the Lasso. Numerical experiments show that this method can outperform direct estimation by the Lasso.

In collaboration with Professor Abdallah Mkhadri (University of Marrakesh, Marocco), Gilles Celeux supervised the thesis of Mohammed El Anbari which concern regularisation methods in linear regression. In collaboration with Professor Abdallah Mkhadri (University of Marrakesh, Marocco), Mohammed El Anbari proposed a method to simultaneously select variables and favor a grouping effect where strongly correlated predictors tend to be in or out of the model together. Numerical experiments showed that their method can be preferred to Elastic-Net when the number of variables is less or equal to the sample size and remain competitive otherwise. Moreover, they have proposed AdaGril an extension of the the adaptive Elastic Net which incorporates information redundancy among correlated variables for model selection and estimation. Under weak conditions, They have established an oracle property of AdaGril. Numerical experiments show in some cases of AdaGril outperforms its competitors.

In collaboration with Jean-Michel Marin (Université de Montpellier) and Christian P. Robert (CEREMADE, Université Paris Dauphine) Gilles Celeux and Mohammed El Anbari highlight the interest of Bayesian regularization methods, using hierarchical non informative priors, compared with standard regularization methods in a poorly informative context through numerical experiments [47].

Clément Levrard worked on the obtention of fast rates of convergence for vector quantization. Using theoretical analogies between quantization seen as an unsupervised learning probel and the one of supervised learning by empirical contrast minimzation, he has obtained a logarithmic improvement on the previously obtained bound. He has been furthermore able to define intellegible "margin type" condition under which fast rates can be obtained.

Since September 2008, Pascal Massart is the cosupervisor with Frédéric Chazal (GEOMETRICA) of the thesis of Claire Caillerie (GEOMETRICA). The project intends to explore and to develop new researches at the crossing of information geometry, computational geometry and statistics.

6.2. Statistical learning methodology and theory

Participants: Gilles Celeux, Christine Keribin, Erwan Le Pennec, Pascal Massart, Lucie Montuelle, Jean-Michel Poggi.

Unsupervised segmentation is an issue similar to unsupervised classification with an added spatial aspect. Functional data is acquired on points in a spatial domain and the goal is to segment the domain in homogeneous domain. The range of applications includes hyperspectral images in conservation sciences, fMRi data and all spatialized functional data. Erwan Le Pennec and Lucie Montuelle are focusing on the questions of the way to handle the spatial component from both the theoretical and the practical point of views as well as the choice of the number of clusters. Furthermore, as functional data require heavy computation, they are required to propose numerically efficient algorithms.

Gilles Celeux, Christine Keribin and the Ph D. student Vincent Brault continue their work on the Latent Block Model. They have proposed an efficient algorithm coupling a Stochastic version of the EM algorithm including a Gibbs sampling step and the Variational EM algorithm. This SEM-VEM algorithm is insensible to its initial position. On the other hand they got a closed formed expression of the Integrated Completed Likelihood for binary tables which allows for a reliable model selection criterion avoiding asymptotic approximation. Moreover, Christine Keribin derived sufficient conditions ensuring the identifiability of the Latent Block Model.

6.3. Reliability and Computer Experiments

Participants: Yves Auffray, Gilles Celeux, Rémi Fouchereau, Shuai Fu, Pascal Massart.

In the computer experiments field, the goal is to approximate an expensive black box function from a limited number of evaluations. The choice of these evaluations i.e. the choice of a design of (computer) experiments is a major issue.

This year Yves Auffray and Pierre Barbillon, in collaboration with Jean-Michel Marin (Université de Montpellier) have considered estimating the probability of rare events in the context of computer experiments. These rare events depends on the output of a physical model with random input variables. Since the model is only known through an expensive black box function, a crude Monte Carlo estimator does not perform well. Two strategies have been developed to cope with this difficulty: a Bayesian estimate and an importance sampling method. Both methods relies on Kriging metamodeling. They are able to achieve sharp upper confidence bounds on the rare event probability. These methods have been applied to a toy example and a real case study which consists of finding an upper bound of the probability that the trajectory of an airborne load collides the aircraft that has released it.

Following the previous work of the first year, Shuai Fu, under the direction of Gilles Celeux, focus on the design of experiments and its validation, which has become the main issues of the thesis. It leads both to theoretical and computational developments. An original DAC criterion has been proposed and leads to a Bayesian procedure of DAC-test to measure the quality of a design. For improving the design of experiments, an adaptive kriging procedure well adapted to the specific problem has been proposed. However, the algorithms require a too important computation time which should be reduced in future work.

In the framework of a CIFRE convention with Snecma-SAFRAN Rémy Fouchereau has started a thesis on the modeling of fatigue damage for Inco718 supervised by Gilles Celeux. Inco718 is a Zinc-based alloy. To determine its minimum lifetime, a lot of stress tests are made. The lloay lifetimes are reported as function of the stress. The aim of this work is to analyse the resulting curves. A mixture model with a lognormal component and a sum of two lognormals components is considered. Since the sum of two or more lognormal distribution is not closed form. Inference on this model needs Monte Carlo integration within the EM algorithm. Despite some unstability for small sample sizes, this model show encouraging and easily interpretable results.

6.4. Statistical analysis of genomic data

Participants: Gilles Celeux, Andrea Rau.

Andrea Rau and Gilles Celeux, in collaboration with Marie-Laure Martin-Magniette (URGV and UMR AgroParisTech/INRA MIA 518) and Cathy Maugis-Rabusseau (IMT/INSA Toulouse) have developed a method to cluster digital gene expression observations from high-throughput (HTS) data using Poisson mixture models [44]. The proposed model has the advantage of accounting for the particularities of HTS data and providing straightforward procedures for parameter estimation and model selection. A series of simulation experiments was done to compare the performance of the proposed model to that of previously proposed clustering methods for similar sequence-based data, and the performance of the proposed approach was examined on two real high-throughput sequencing data sets. The R package HTSCluster used to implement the proposed Poisson mixture model has been made freely available on CRAN.

6.5. Curves classification, denoising and forecasting

Participant: Jean-Michel Poggi.

In collaboration with Farouk Mhamdi and Meriem Jaidane (ENIT, Tunis, Tunisia), Jean-Michel Poggi proposed, in [18]. a method for trend extraction from seasonal time series through the Empirical Mode Decomposition (EMD). Experimental comparison of trend extraction based on EMD, X11, X12 and Hodrick Prescott filter are conducted. First results show the eligibility of the blind EMD trend extraction method. Tunisian real peak load is also used to illustrate the extraction of the intrinsic trend.

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 [3].

Jean-Michel Poggi was the supervisor (with A. Antoniadis) of the PhD Thesis of Jairo Cugliari-Duhalde which takes place in a CIFRE convention with EDF. It is strongly related to the use of wavelets together with curves clustering in order to perform accurate load comsumption forecasting. The thesis develops methodological and applied aspects linked to the electrical context as well as theoretical ones by introducing exogeneous variables in the context of nonparametric forecasting time series (see [27] and [45]).

6.6. Neuroimaging, Statistical analysis of fMRI data

Participants: Gilles Celeux, Christine Keribin.

This research takes place as part of a collaboration with Neurospin on brain functional Magnetic Resonance Imaging (fMRI) data. (http://www.math.u-psud.fr/select/reunions/neurospin/Welcome.html). This year it concerned essentially regularisation in a supervised clustering methodology that includes spatial information in the prediction framework, and yields clustered weighted maps.

7. Contracts and Grants with Industry

7.1. Contracts with EDF

Participants: Gilles Celeux, Jean-Michel Poggi.

- SELECT has a contract with EDF regarding modelling uncertainty in deterministic models.
- SELECT has a contract with EDF regarding wavelet analysis of the electrical load consumption for the aggregation and desaggregation of curves to improve total signal prediction.

7.2. Other contracts

Participants: Gilles Celeux, Rémy Fouchereau.

- SELECT has a contract with SAFRAN - SNECMA, an high-technology group (Aerospace propulsion, Aicraft equipment, Defense Security, Communications),regarding modelling reliability of Aircraft Equipment (collaboration with Patrick Pamphile (Université Paris-Sud).

8. Partnerships and Cooperations

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 is co-organizing a working group at ENS (Ulm) on Statistical Learning. This year the group focused interest on regularization 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.

SELECT is animating a working group on Unsupervised Classification with the CMAP (École Polytechnique)

8.2. European Initiatives

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

8.3. International Initiatives

Gilles Celeux is one of the co-organizers of the Working Group on Model-Based Clustering.

9. Dissemination

9.1. Scientific Community 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 *CSBIGS* and *La Revue Modulad*.
- Pascal Massart is Associated Editor of *Annals of Statistics*, *Confluentes Mathematici*, and *Foundations and Trends in Machine Learning*.
- Jean-Michel Poggi is Associated Editor of *Journal of Statistical Software*, *Journal de la SFdS* and *CSBIGS*.

9.1.2. Invited conferences

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

- Gilles Celeux was invited speaker to IFCS 2011 in Frankfurt, to the mixture session of JSM2011 in Miami, to StatSeq 2011 in Toulouse, to the statistical seminar of the Economics departement of Vienna University and to the Summer Model-Based Clustering working group in Glasgow.
- Jean-Michel Poggi was invited speaker at SIS 2011, 46th Scient. Meeting of the Italian Stat. Society in Bologna, at ENBIS-11 in Coimbra and at the Worksoph - In honour of Anestis Antoniadis at Villard de Lans.

9.1.3. Scientific animation

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

- Gilles Celeux is member of the CSS of INRA.
- Gilles Celeux was Chair of the Chikio Hayashi Awards Committee.
- Erwan Le Pennec is a member of the Board of the MAS group of the SMAI (french SIAM).
- Erwan Le Pennec and Pascal Massart are members of the C.N.U. (section 26).
- Pascal Massart is a senior member of the I.U.F.
- Pascal Massart is a member of the scientific council of the French Mathematical Society.
- Pascal Massart is a member of the scientific council of the Mathematical Department of the Ecole Normale Supérieure de Paris.
- Pascal Massart was a member of the scientific committee of the European Meeting of Staticians in Piraeus.
- Jean-Michel Poggi is Cochair seminar of Probability and 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 Chair of the Program Commitee of the «Journées de Statistique de la SFdS», Tunis, mai 2011
- Jean-Michel Poggi is President of the French statistical society (SFdS).
- Jean-Michel Poggi is member of the Board of the "Environment group" of the French statistical society (SFdS).

9.2. Teaching

Master: Gilles Celeux, modèles à structure cachée ISUP 3ème année (Université Paris 6) 20 heures

Master: Gilles Celeux, modèles pour la classification M2 probabilités et statistique, Université Paris Sud, 24 heures

Master: Erwan Le Pennec, Méthodes d'ondelettes, 24h, Mé, Université Paris Diderot, France

Master: Erwan Le Pennec, Analyse Spectrale, 18h, M1, Ponts Paristech, France

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

PhD & HdR :

PhD : Jairo Cugliari Duhalde, Prévision d'un processus à valeurs fonctionnelles. Application à la consommation d'électricité, 22/11/2011 at Paris XI Orsay, J.-M. Poggi and Anestis Antoniadis (Univ. Joseph Fourier, Grenoble)

PhD : Robin Genuer, Forêts aléatoires : aspects théoriques, sélection de variables et applications, 24/11/2010 at Paris XI Orsay, J.-M. Poggi

PhD in progress: Vincent Brault, 2011, Gille Celeux and Christine Keribin

PhD in progress: Claire Caillerie, 2008, Pascal Massart and Frédéric Chazal

PhD in progress: Rémi Fouchereau, 2011, Gille Celeux

PhD in progress: Shuai Fu, 2010, Gille Celeux

PhD in progress: Clément Levrard, 2009, Pascal Massart and Gérard Biau (UPMC)

PhD in progress: Caroline Meynet, 2009, Pascal Massart

PhD in progress: Lucie Montuelle, Sélection de modèles et mélange de gaussiennes en imagerie hyperspectrale, 01/10/2011, Erwan Le Pennec

PhD in progress: Nelo Molter Magalães, 2011, Pascal Massart

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] J. C. DUHALDE. *Prévision d'un processus à valeurs fonctionnelles. Application à la consommation d'électricité*, Paris Sud, Orsay, 2011.

[2] R. GENUER. *Forêts aléatoires : aspects théoriques, sélection de variables et applications*, Paris Sud, Orsay, 2011.

Articles in International Peer-Reviewed Journal

[3] M. AMINGHAFARI, J.-M. POGGI. *Multistep Forecasting Non-Stationary Time Series using Wavelets and Kernel Smoothing*, in "Communications in Statistics, Theory and Methods", 2011, to appear.

[4] M. AMINGHAFARI, J.-M. POGGI. *Multistep Forecasting Non-Stationary Time Series using Wavelets and Kernel Smoothing*, in "Communications in Statistics Theory and Methods", 2012, vol. 41, p. 1–15.

[5] Y. Auffray, P. Barbillon, J.-M. Marin. *Maximin Design on non-hypercube domain and Kernel Interpolation*, in "Statistics and Computing", 2011 [*DOI :* 10.1007/s11222-011-9273-9], http://hal.inria.fr/inria-00638728/en.

[6] P. Barbillon, G. Celeux, A. Grimaud, Y. Lefebvre, E. De Rocquigny. *Non linear methods for inverse statistical problems*, in "Computational Statistics and Data Analysis", 2011 [*DOI :* 10.1016/j.csda.2010.05.030], http://hal.inria.fr/inria-00441967/en.

[7] J.-P. Baudry, C. Maugis, B. Michel. *Slope heuristics: overview and implementation*, in "Statistics and Computing", 2011, vol. 22, p. 455-470.

[8] K. Bertin, E. Le Pennec, V. Rivoirard. *Adaptive Dantzig density estimation*, in "Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques", 2011, vol. 47, n^o 1, p. 43-74 [*DOI :* 10.1214/09-AIHP351], http://hal.inria.fr/hal-00381984/en.

[9] L. Bertrand, M.-A. Languille, S. Cohen, L. Robinet, C. Gervais, S. Leroy, D. Bernard, E. Le Pennec, W. Josse, J. Doucet, S. Schöder. *European research platform IPANEMA at the SOLEIL synchrotron for ancient and historical materials*, in "Journal of Synchrotron Radiation", 2011, vol. 18, n^o 5, p. 765-772 [*DOI :* 10.1107/S090904951102334X], http://hal.inria.fr/hal-00618143/en.

[10] M. Bobbia, F.-X. Jollois, J.-M. Poggi, B. Portier. *Quantifying local and background contributions to PM10 concentrations in Haute-Normandie, using random forests*, in "Environmetrics", 2011, vol. 22, n^o 6, p. 758–768.

[11] S. Boucheron, P. Massart. *A high-dimensional Wilks phenomenon*, in "Probability Theory and Related Fields", January 2011, vol. 150, p. 405-433, http://dx.doi.org/10.1007/s00440-010-0278-7.

[12] C. Bouveyron, G. Celeux, S. Girard. *Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA*, in "Pattern Recognition Letters", 2011, vol. 32, p. 1706-1713 [*DOI :* 10.1016/j.patrec.2011.07.017], http://hal.inria.fr/hal-00440372/en.

[13] G. Celeux, M.-L. Martin-Magniette, C. Maugis, R. Adrian. *Letter to the Editor*, in "Journal of the American Statistical Association", 2011, vol. 106, p. 383-383.

[14] C. Dossal, E. Le Pennec, S. Mallat. *Bandlet Image Estimation with Model Selection*, in "Signal Processing", January 2011, vol. 91, n^o 12, p. 2743-2753 [*DOI :* 10.1016/j.sigpro.2011.01.013], http://hal.inria.fr/hal-00321965/en.

[15] M. Lavielle, A. Samson, A. Karina Fermin, F. Mentré. *Maximum likelihood estimation of long-term HIV dynamic models and antiviral response.*, in "Biometrics", March 2011, vol. 67, n^o 1, p. 250-9 [*DOI :* 10.1111/j.1541-0420.2010.01422.x], http://hal.inria.fr/inserm-00486937/en.

[16] P. Massart, C. Meynet. *An ℓ_1-oracle inequality for the Lasso*, in "Electronic Journal of Statistics", 2011, vol. 5, p. 669–687, http://hal.inria.fr/inria-00506446/PDF/RR-7356.pdf.

[17] C. Maugis, G. Celeux, M.-L. Martin-Magniette. *Variable selection in Model-based iscriminant Analysis*, in "Journal of Multivariate Analysis", 2011, vol. 102, p. 1374-1387.

[18] F. MHAMDI, J.-M. POGGI, M. JAIDANE. *Trend Extraction for Seasonal Time Series using Ensemble Empirical Mode Decomposition*, in "Advances in Adaptive Data Analysis (AADA)", 2011, vol. 3, n^{o} 3, p. 363–383.

[19] V. MICHEL, E. EGER, C. KERIBIN, B. THIRION. *Multiclass Sparse Bayesian Regression for fMRI-Based Prediction*, in "International Journal of Biomedical Imaging", April 2011, vol. 2011, epub [*DOI :* 10.1155/2011/350838], http://hal.inria.fr/inria-00609365/en.

[20] V. MICHEL, A. GRAMFORT, G. VAROQUAUX, E. EGER, C. KERIBIN, B. THIRION. *A supervised clustering approach for fMRI-based inference of brain states*, in "Pattern Recognition", April 2011, epub ahead of print [*DOI :* 10.1016/J.PATCOG.2011.04.006], http://hal.inria.fr/inria-00589201/en.

[21] V. MICHEL, A. GRAMFORT, G. VAROQUAUX, E. EGER, C. KERIBIN, B. THIRION. *A supervised clustering approach for fMRI-based inference of brain states*, in "Pattern Recognition - Special Issue on Brain Decoding", 2011, epub ahead of print.

[22] A. PASANISI, S. FU, N. BOUSQUET. *Estimating discrete Markov models from various incomplete data schemes*, in "Computational Statistics & Data Analysis", 2011, to appear.

[23] J.-M. POGGI, B. PORTIER. *PM10 forecasting using clusterwise regression*, in "Atmospheric Environment", 2011, vol. 45, n^{o} 38, p. 7005–7014.

[24] A. RAU, F. JAFFRÉZIC, J.-L. FOULLEY, R. DOERGE. *Reverse Engineering Gene Networks Using Approximate Bayesian Computation (ABC*, in "Statistics and Computing", 2011, to appear.

Articles in National Peer-Reviewed Journal

[25] Y. AUFFRAY, P. BARBILLON, J.-M. MARIN. *Modèles réduits à partir d'expérience numériques*, in "Journal de la Société Française de Statistique", 2011, http://hal.inria.fr/inria-00638735/en.

[26] C. KERIBIN. *Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie*, in "Journal de la Sociéte Fran çaise de Statistiques", 2011, vol. 151, n^{o} 2, p. 107-131.

Invited Conferences

[27] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Functional Clustering using Wavelets*, in "Invited conference at SIS 2011, 46th Scient. Meeting of the Italian Stat. Society", Bologna (Italy), 8-10 june 2011.

[28] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Functional Clustering using Wavelets*, in "Invited conference at SIS 2011, 46th Scient. Meeting of the Italian Stat. Society", Bologna (Italy), 8-10 june 2011.

[29] G. CELEUX. *Model -based cluster analysis for transcriptomic data*, in "StaSeq2011", Toulouse, France, April 2011.

[30] G. CELEUX. *Model Selection for the Latent Block Model*, in "Working group on Model-Based Clustering", Glasgow, Scotland, July 2011.

[31] G. CELEUX. *Recent Advances in Finite Mixture Models and Clustering*, in "Joint Statistical Meeting", Miami, USA, August 2011.

[32] G. CELEUX. *Statistical inference for the latent block model: a review*, in "Symposium of the International Federation of Classification Societies", Frankfurt, Germany, August 2011.

[33] S. FU, G. CELEUX, M. COUPLET, N. BOUSQUET. *A Bayesian solution to characterizing uncertainty in inverse problems*, in "The 58th ISI Congress", Dublin, Ireland, August 2011.

[34] S. FU, G. CELEUX, M. COUPLET, N. BOUSQUET. *A Bayesian solution to characterizing uncertainty in inverse problems*, in "The 3rd IMS-China International conference on Statistics and Probability", Xi'an, China, July 2011.

[35] A. B. HEN, S. GEY, J.-M. POGGI. *Detecting Influent Observations using CART Classification Trees. Application to the classification of the cities of Paris area*, in "ENBIS-11", Coimbra (Portugal), 4-8 september 2011.

[36] A. B. HEN, S. GEY, J.-M. POGGI. *Influence measures for CART Classification Trees*, in "Workshop In honour of Anestis Antoniadis", Villard de Lans, 23-25 march 2011.

National Conferences with Proceeding

[37] S. COHEN, E. LE PENNEC. *Segmentation non supervisée d'image hyperspectrale par mélange de gausiennes spatialisé*, in "GRETSI 2011", Bordeaux, France, September 2011, http://hal.inria.fr/inria-00638432/en.

Conferences without Proceedings

[38] S. FU, G. CELEUX, M. COUPLET, N. BOUSQUET. *A Bayesian solution to characterizing uncertainty in inverse problems*, in "Journées des doctorants, GdR MASCOT-NUM", Villard de Lans, France, March 2011.

Scientific Books (or Scientific Book chapters)

[39] G. CELEUX. *11*, in "Bayesian Inference and Markov Chain Monte Carlo Methods", Wiley, 2011, p. 207-226.

Research Reports

[40] S. COHEN, E. LE PENNEC. *Conditional Density Estimation by Penalized Likelihood Model Selection and Applications*, INRIA, April 2011, n^{o} RR-7596, http://hal.inria.fr/inria-00575462/en.

[41] E. GAUTIER, E. LE PENNEC. *Adaptive estimation in the nonparametric random coefficients binary choice model by needlet thresholding*, INRIA, June 2011, n^{o} RR-7647, http://hal.inria.fr/inria-00601274/en.

[42] R. GENUER, I. MORLAIS, W. TOUSSILE. *Gametocytes infectiousness to mosquitoes: variable selection using random forests, and zero inflated models*, INRIA, January 2011, n^{o} RR-7497, http://hal.inria.fr/inria-00550980/en.

[43] P. MASSART, R. RAPHAEL. *Around Nemirovski's inequality*, Inria, 2011.

[44] A. Rau, G. Celeux, M.-L. Martin-Magniette, C. Maugis-Rabusseau. *Clustering high-throughput sequencing data with Poisson mixture models*, INRIA, November 2011, n^{o} RR-7786, http://hal.inria.fr/inria-00638082/en.

Other Publications

[45] A. Antoniadis, X. Brossat, J. Cugliari, J.-M. Poggi. *Functional Clustering using Wavelets*, 2011, Preprint HAL.

[46] Y. Auffray, P. Barbillon, J.-M. Marin. *Estimation of rare events probabilities in computer experiments*, 2011, 20 pages, 6 figures, http://hal.inria.fr/inria-00638696/en.

[47] G. Celeux, M. E. Anbari, J.-M. Marin, C. P. Robert. *Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation*, 2011, To appear, http://hal.archives-ouvertes.fr/hal-00523354/en/.

[48] A. B. Hen, S. Gey, J.-M. Poggi. *Influence functions for CART*, 2011, Preprint HAL.

[49] A. B. Hen, S. Gey, J.-M. Poggi. *Influence measures for CART Classification Trees*, 2011, Preprint HAL.