IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2012

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

Hébergé au Laboratoire de Mathématiques d'Orsay, Faculté des sciences, Université Paris-Sud

Creation of the Project-Team: January 01, 2007 .

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]
Patrick Pamphile [Associate Professor]
Jean-Michel Poggi [Professor UniversitéParis 5, HdR]

External Collaborators

Yves Auffray [Dassault]
Serge Cohen [CNRS - Synchrotron Soleil]
Michel Prenat [Thales Optronique]

PhD Students

Vincent Brault [MESR grant]
Mohammed El Anbari [France-Marocco grant]
Émilie Devijver [MESR grant]
Rémy Fouchereau [MESR grant]
Shuai Fu [EDF-Inria Cifre grant]
Mélina Galopin [MESR grant]
Clément Levrard [MESR grant]
Caroline Meynet [MESR grant]
Nelo Molter Magalães [MESR grant]
Lucie Montuelle [MESR grant]
Vincent Thouvenot [EDF Cifre Grant]
Solenne Thivin [Thales Optronic Cifre grant]

Post-Doctoral Fellows

Mesrob Ohannessian
Adrien Saumard [Fondation grant]
Mohammed Sedki [Inria Grant]
Tim van Erven [Rubicon Grant]
Jairo Cugliari [Inria]

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, neuroimaging and ancient material imaging.

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, Michel Misiti, Yves Misiti and Bruno Portier (INSA de Rouen), in the context of a collaboration between AirNormand, Orsay University and INSA of Rouen. Mixtures of linear regression models are used for the short-term statistical forecasting of the daily mean PM10 concentration in three cities in Haute-Normandie (France): Rouen, Le Havre and Dieppe. The Haute-Normandie region is located at northwest of Paris, near the south side of Manche sea and is heavily industrialized. Six monitoring stations reflecting the diversity of situations: urban background, traffic, rural and industrial stations are considered. This recent statistical method has been used and beyond the application, this study shed light on this method [35].

4.7. Analysis spectroscopic imaging of ancient materials

Ancient materials, encountered in archaeology, paleontology and cultural heritage, are often complex, heterogeneous and poorly characterised before their physico-chemical analysis. A technique of choice to gather as much physico-chemical information as possible is spectro-microscopy or spectral imaging where a full spectra, made of more than thousand samples, is measured for each pixel. The produced data is tensorial with two or three spatial dimensions and one or more spectral dimensions and it requires the combination of an «image» approach with «curve analysis» approach. Since 2010 SELECT collaborates with Serge Cohen (IPANEMA) on the development of conditional density estimation through GMM and non-asymptotic model selection to perform stochastic segmentation of such tensorial dataset. This technic enablesx the simultaneous accounting for spatial and spectral information while producing statistically sound information on morphological and physico-chemical aspects of the studied samples.

5. Software

5.1. MIXMOD software

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

Mixture model, cluster analysis, discriminant analysis

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 2010, 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 now available http://cran.r-project.org/web/packages/Rmixmod/index.html.

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, Erwan Le Pennec, Lucie Montuelle, Pascal Massart, Caroline Meynet, Jean-Michel Poggi, Adrien Saumard.

Erwan Le Pennec is still working with Serge Cohen (IPANEMA Soleil) on hyperspectral image segmentation based on a spatialized Gaussian Mixture Model. Their scheme is supported by some theoretical investigation [6] and have been applied in pratice with an efficient minimization algorithm combining EM algorithm, dynamic programming and model selection implemented with MIXMOD. Lucie Montuelle is studying extensions of this model that comprise parametric logistic weights and regression mixtures.

In collaboration with Marie-Laure Martin-Magniette (URGV et UMR AgroParisTech/INRA MIA 518) and Cathy Maugis (INSA Toulouse) Gilles Celeux 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. Moreover they have analysed the differences between the model-based approach and geometrical approach to select variable for clustering. Through numerical experiments, they showed the advantage of the model-based appraoch when many variables are highly correlated. 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 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 Jean-Patrick Baudry (Paris 6) and Margarida Cardoso, Ana Ferreira and Maria-José Amorim (Lisbon University], Gilles Celeux has proposed an approach to select, in the model-based clustering context, a model and a number of clusters in order to get a partition which both provides a good fit with the data and is related to the external categorical variables. This approach makes use of the integrated joint likelihood of the data, the partition derived from the mixture model and the known partitions. It is worth noticing that the external categorical variables are only used to select a relevant mixture model. Each mixture model is fitted by the maximum likelihood methodology from the observed data. Numerical experiments illustrate the promising behaviour of the derived criterion [29].

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.

Tim van Erven is studying Model Selection for the Long Term. When a model selection procedure forms an integrated part of a company's day-to-day activities, its performance should be measured not on a single day, but on average over a longer period, like for example a year. Taking this long-term perspective, it is possible to aggregate model predictions optimally even when the data probability distribution is so irregular that no statistical guarantees can be given for any individual day seperately. He studies the relation between model selection for individual days and for the long term, and how the geometry of the models affects both. This work has potential applications in model aggregation for the forecasting of electrical load consumption at EDF.

Adrien saumard has worked on the theoretical validation of the slope heuristics, a practical method of penalties calibration derived in a Gaussian setting by Birgé and Massart in 2006 and extended to bounded M-estimation by Arlot and Massart in 2010. He was able to prove the validity of this heuristics in bounded heteroscedastic regression with random design when the considered models where linear spans made of

piecewise polynomials. A preliminary work on a fixed model was necessary and published in [9], while the validation of the slope heuristics itself - as well as the validation of a cross-validation approach - can be found in a preprint.

6.2. Statistical learning methodology and theory

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

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. They study in particular the choice of the number of clusters. Furthermore, as functional data require heavy computation, they are required to propose numerically efficient algorithms. They have also extend the model to regression mixture.

Gilles Celeux, Christine Keribin and the Ph D. student Vincent Brault continue their work on the Latent Block Model (LBM). They compared several model selection criteria for binary tables [19]. However, the SEM-VEM Gibbs algorithm used to estimate LBM is subject to spurious solutions (empty clusters). To tackle this drawback, they have proposed to use Bayesian inference through Gibbs Sampling and studied the influence of the calibration of non informative prior distributions. They showed on numerical experiment the advantages of coupling Gibbs sampling with a Variational Bayes algorithm to get pointwise estimators [17]. Furthermore, they extended the previous studies from binary to categorical data [32].

Christine Keribin has proposed to compare, on genomics applications, the use of LBM with other methodologies (variable selection procedure of Maugis and Martin Magniette, component analysis). She supervised an internship (Master 1) on the use of principal component analysis for gene expression data (Inria funding). This has been done on data of the SONATA project (leaded by URGV - Evry Genopole), in collaboration with Marie-Laure Martin-Magniette.

Erwan Le Pennec is supervising Solenne Thivin in her CIFRE with Michel Prenat and Thales Optronique. The aim is target detection on complex background such as clouds or sea. Their approach is a local test approach based on the test decision theory. A key issue is to learn good discrimant features and their probabilistic properties. So far, they have worked on cloud images given by Thales. They focus on a Markovian modeling of the clouds.

Considering the case of maximum likelihood density estimation on histograms, Adrien saumard has investigated both theory and methodology. On the one hand, he has shown that AIC is twice the minimal penalty in the sense of Birgé and Massart, which by consequence implies the asymptotic optimality of the slope heuristics based on a linear shape. On the other hand, he investigated the methodology of the small to moderate sample size setting in this case. The robustness of the slope heurisitics compared to AIC is shown on simulated examples and a new overpenalization of Akaike's criterion is proposed, which outperforms the criterion AICc of Hurvitch and Tsai and shows comparable results to the procedure proposed by Birgé and Rozenholc in 2006. The benefits of the derived procedure here is its theoretical background and interpretation. This work is still in process and some of the results can be found in a preprint.

6.3. Reliability and Computer Experiments

Participants: Yves Auffray, Gilles Celeux, Rémy Fouchereau, Shuai Fu.

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.

Following the previous work of the past three years, Shuai Fu has concluded her Ph.D thesis under the direction of Gilles Celeux [1]. This year, the work was focused on controlling four main error quantities, in order to validate the methodology in the industrial framework. More precisely, the DAC criterion (Data Agreement Criterion), which has been proposed for assessing the relevance of the design of experiments (DOE) and the prior choice with the observed data was applied to a complex hydrological model, coding and testing the relevant algorithms [30]. For the purpose of controlling the emulator error in an adaptive kriging algorithm, two Bayesian criteria have been proposed for searching and adding new points into the current DOE. The computation time remains important, which makes the method meaningful only in the case where we have a really time-consuming code.

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 alloy lifetimes are reported as function of the stress. The aim is to propose a stochastic models for fatigue lifetime prediction based on a fracture mechanics-based approach. 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. Thus, we have provided engineers with a probabilistic tool for reliability design of mechanical parts, but also with a diagnostic tool for material elaboration.

6.4. Statistical analysis of genomic data

Participant: Gilles Celeux.

In collaboration with Florence Jaffrezic and Andrea Rau (INRA, département de génétique animale) Gilles Celeux initiated modelling genomics networks from RNA-seq data. It was the subject of the internship of Mélina Gallpin who is starting a thesis on this subject. To day the performance of overdispersed Poisson models has been investigated. The results are somewhat poor especially for large numbers of genes.

6.5. Curves classification, denoising and forecasting

Participants: Émilie Devijver, Pascal Massart, Jean-Michel Poggi.

In collaboration with Farouk Mhamdi and Meriem Jaidane (ENIT, Tunis, Tunisia), Jean-Michel Poggi proposeda 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.

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.

Jean-Michel Poggi, co-supervising with Anestis Antoniadis (Université Joseph Fourier Grenoble) the PhD thesis of Vincent Thouvenot, funded by a CIFRE with EDF. The industrial motivation of this work is the recent development of new technologies for measuring power consumption by EDF to acquire consumption data for different mesh network. The thesis will focus on the development of new statistical methods for predicting power consumption by exploiting the different levels of aggregation of network data collection.

From the mathematical point of view, the work is to develop generalized additive models for this type of kind of aggregated data for the modeling of functional data, associating closely nonparametric estimation and variable selection using various penalization methods.

Jean-Michel Poggi and Pascal Massart are the co-advisors of the PhD thesis of Emilie Devijver, strongly motivated by the same kind of industrial forecasting problems in electricity, is dedicated to curves clustering for the prediction. A natural framework to explore this question is mixture of regression models for functional data. The theoretical subject of the thesis is to extend to functional data the recent work by Bühlmann et al. dealing with the simultaneous estimation of mixture regression models in the scalar case using Lasso type methods. Of course, it will be based on the technical tools of the work of Caroline Meynet (which completes his thesis Orsay under the direction of P. Massart), which deals with the clustering of functional data using Lasso methods choosing simultaneously number of clusters and selecting significant wavelet coefficients.

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). and concerns essentially regularisation in a supervised clustering methodology that includes spatial information in the prediction framework, and yields clustered weighted maps.

7. Bilateral 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, Patrick Pamphile.

- 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. Regional Initiatives

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. This year this workshop took place in Guelph (Canada).

9. Dissemination

9.1. Scientific Animation

9.1.1. Editorial responsibilities

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

- Gilles Celeux is Editor-in-Chief of *Statistics and Computing* and of *Journal de la SFdS*. 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 the Method and Research meetings at UNIL (Lausanne) and to the Summer Model-Based Clustering working group in Guelp.
- Pascal Massart has given the "Le Cam lecture" at the last worldwide IMS-Bernoulli meeting in Istambul.
- Jean-Michel Poggi was invited speaker at SIS 2012, and at 5th International Conference of the ERCIM Working Group on Computing and Statistics, Oviedo, Spain.

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 a member of the scientific committee of SMPGD (Statistical Methods for Post Genomics Data).
- Erwan Le Pennec is a member of the Board of the MAS group of the SMAI (french SIAM).
- Erwan Le Pennec is a member of the Labex AMIES (Agence pour les Mathématiques en Interaction avec les Entreprises et la Société).
- 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 President of the French statistical society (SFdS).
- Jean-Michel Poggi is Vice-President of FENStatS «Federation of European National Statistical Societies»
- Jean-Michel Poggi is Member of the Program Commitee of WIPFOR "Workshop on Industry & Practices for Forecasting", June 5-7, 2013, Paris

9.2. Teaching - Supervision - Juries

9.2.1. 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, Méthode Parcimonieuse en Statistique, 30h, Université Paris Sud, France

Master: Erwan Le Pennec, Méthodes d'ondelettes, 20h, M2, Université Paris Diderot, France

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

Master: Jean-Michel Poggi, Ondelettes et applications (Master 2 Ingénierie Mathématique, Université Paris Sud) 30 heures

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.

9.2.2. Supervision

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: Caroline Meynet, 2009, Pascal Massart

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: Émilie Devivjer, 2012, Pascal Massart and Jean-Michel Poggi

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

PhD in progress: Farouk Mhamdi, 2012, Jean-Michel Poggi and Meriem Jaïdane (ENIT Tunisie)

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

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

PhD in progress: Solenne Thivin, 2012, Erwan Le Pennec

PhD in progress: Vincent Thouvenot, 2012, Jean-Michel Poggi and Anestis Antoniadis (Univ. Joseph Fourier, Grenoble)

9.3. Popularization

Erwan Le Pennec takes care of a Math en Jeans group at lycée Joliot Curie from Nanterre.

Vincent Brault takes care of a Math en Jeans group at lycée Blaise Pascal from Orsay and is part of the organizing committee of the Orsay conference for April 2013

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] S. FU. *Inversion probabiliste bayésienne en analyse d'incertitude*, Université Paris-sud 11, 2012.

[2] C. MEYNET. *Sélection de variables pour la classification non supervisée en grande dimension*, Université Paris-sud 11, 2012.

Articles in International Peer-Reviewed Journals

[3] 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, n^{o} 3, p. 485–499.

[4] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Functional Clustering using Wavelets*, in "International Journal of Wavelets, Multiresolution and Information Processing", 2012, Accepted for publication.

[5] J.-P. BAUDRY, C. MAUGIS, B. MICHEL. *Slope heuristics: overview and implementation*, in "Statistics and Computing", 2012, vol. 22, n^{o} 2, p. 455-470 [*DOI :* 10.1007/s11222-011-9236-1], http://hal.inria.fr/hal-00666838.

[6] S. X. COHEN, E. LE PENNEC. *Partition-Based Conditional Density Estimation*, in "ESAIM: Probability and Statistics", 2012 [*DOI :* 10.1051/ps/2012017], http://hal.inria.fr/hal-00752943.

[7] G. KERKYACHARIAN, E. LE PENNEC, D. PICARD. *Radon needlet thresholding*, in "Bernoulli", 2012, vol. 18, n^{o} 2, p. 391-433 [*DOI :* 10.3150/10-BEJ340], http://hal.inria.fr/hal-00409903.

[8] P. MASSART, C. MEYNET. *Around Nemirovski's inequality*, in "IMS collections", 2012, vol. 9, p. 254–265.

[9] A. SAUMARD. *Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression*, in "Electron. J. Statist.", 2012, vol. 6, n^{o} 1-2, p. 579–655.

[10] T. VAN ERVEN, M. REID, R. WILLIAMSON. *Mixability is Bayes Risk Curvature Relative to Log Loss*, in "Journal of Machine Learning Research, special issue on Inductive Logic Programming", May 2012, n^{o} 13, p. 1639–1663, http://hal.inria.fr/hal-00758204.

[11] V. VANDEWALLE, C. BIERNACKI, G. CELEUX, G. GOVAERT. *A predictive deviance criterion for selecting a generative model in semi-supervised classification*, in "Computational Statistics and Data Analysis", 2012, to appear.

Articles in National Peer-Reviewed Journals

[12] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. Application à la consommation d'électricité*, in "Journal de la Société Française de Statistique", 2012, Accepted for publication.

International Conferences with Proceedings

[13] V. BRAULT, J.-P. BAUDRY, C. MAUGIS-RABUSSEAU, B. MICHEL. *Package Capushe pour le logiciel R*, in "44ème journées de statistique", Université libre de Bruxelles, campus du Solbosch, 05 2012, http://jds2012.ulb.ac.be/myreview/files/default/submission/submission_184.pdf.

[14] V. BRAULT, J.-P. BAUDRY, C. MAUGIS-RABUSSEAU, B. MICHEL. *Package Capushe pour le logiciel R*, in "1ère Rencontres R", Université de Bordeaux, campus Victoire, 07 2012, http://hal.archives-ouvertes.fr/hal-00717565.

[15] V. BRAULT, G. CELEUX, C. KERIBIN. *Régularisation bayésienne du modèle des blocs latents*, in "44ème journées de statistique", Université libre de Bruxelles, campus du Solbosch, 05 2012, http://jds2012.ulb.ac.be/myreview/files/default/submission/submission_126.pdf.

[16] V. BRAULT, G. CELEUX, C. KERIBIN. *Régularisation bayésienne du modèle des blocs latents*, in "Workshop ClasSel", IHP, 11 rue Pierre et Marie Curie - 75231 Paris, 01 2012, https://sites.google.com/site/workshopclassel/resumes.

[17] V. BRAULT, G. CELEUX, C. KERIBIN. *Régularisation bayésienne du modèle des blocs latents*, in "44èmes Journées de Statistique, Bruxelles, France", 2012, http://jds2012.ulb.ac.be/myreview/files/default/submission/submission_126.pdf.

[18] R. FOURCHEREAU, G. CELEUX, P. PAMPHILE. *Probabilistic modelling of SN curve*, in "S2MRSA", Bordeaux, France, July 4-6th 2012.

[19] C. KERIBIN, V. BRAULT, G. CELEUX, G. GOVAERT. *Model selection for the binary latent block model*, in " 20th International Conference on Computational Statistics (COMPSTAT 2012", Limassol, Cyprus, August 2012.

[20] P. MASSART, C. MEYNET. *Some Rates of Convergence for the Selected Lasso Estimator*, in "23rd International Conference Algorithmic Learning Theory 2012", Lyon, France, Springer Berlin/Heidelberg, October 29-31 2012, p. 17–33, Algorithmic Learning Theory.

[21] F. MHAMDI, M. JAIDANE, J.-M. POGGI. *Forecasting time series through reconstructed multiple seasonal patterns using Empirical Mode Decomposition*, in "Proceedings of the 20th International Conference on Computational Statistics COMPSTAT2012", Limassol, Cyprus, 2012, p. 573–583.

[22] T. VAN ERVEN, P. GRÜNWALD, M. REID, R. WILLIAMSON. *Mixability in Statistical Learning*, in "Advances in Neural Information Processing Systems 25 (NIPS 2012)", Lake Tahoe, United States, December 2012.

National Conferences with Proceeding

[23] R. FOURCHEREAU, G. CELEUX, P. PAMPHILE. *Modélisation Statistique des données de fatigue matériau*, in "Actes du congrès lambdamu18-IMDR", 2012.

Conferences without Proceedings

[24] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Clustering functional data using wavelets*, in "8th World Congress Proba & Stat", Istanbul, 2012.

[25] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Non parametric forecasting of a function-valued non stationary processes. Application to the electricity demand*, in "5th International Conference of the ERCIM Working Group on Computing and Statistics", Oviedo, Spain, 2012.

[26] M. MISITI, Y. MISITI, J.-M. POGGI, B. PORTIER. *PM10 forecasting using mixture linear regression models*, in "46th scientific meeting of the Italian Statistical Society, SIS 2012", Rome, 2012.

[27] M. MISITI, Y. MISITI, J.-M. POGGI, B. PORTIER. *PM10 forecasting using mixture linear regression models*, in "ENBIS 2012", Ljubana, 2012.

Research Reports

[28] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. Application à la consommation d'électricité.*, Inria, June 2012, n° RR-7982, http://hal.inria.fr/hal-00703570.

[29] J.-P. BAUDRY, M. CARDOSO, G. CELEUX, M.-J. AMORIM, A. SOUSA FERREIRA. *Enhancing the selection of a model-based clustering with external qualitative variables*, Inria, October 2012, n° RR-8124, 14, http://hal.inria.fr/hal-00747387.

[30] S. FU, G. CELEUX, N. BOUSQUET, M. COUPLET. *Bayesian inference for inverse problems occurring in uncertainty analysis*, Inria, June 2012, n° RR-7995, http://hal.inria.fr/hal-00708814.

Other Publications

[31] J.-P. BAUDRY, M. CARDOSO, G. CELEUX, M.-J. AMORIM, A. SOUSA FERREIRA. *Enhancing the selection of a model-based clustering with external qualitative variables*, 2012, HAL, http://hal.inria.fr/hal-00747854.

[32] V. BRAULT, G. CELEUX, G. GOVAERT, C. KERIBIN. *Estimation and Selection for the Latent Block Model on nominal data*, 2012.

[33] S. X. COHEN, E. LE PENNEC. *Conditional Density Estimation by Penalized Likelihood Model Selection*, 2012, Submitted.

[34] C. MEYNET, C. MAUGIS-RABUSSEAU. *A sparse variable selection procedure in model-based clustering*, June 2012, HAL, http://hal.inria.fr/hal-00734316.

[35] M. MISITI, Y. MISITI, J.-M. POGGI, B. PORTIER. *Mixture of linear regression models for short term PM10 forecasting in Haute Normandie (France)*, 2012, Submitted.

[36] T. VAN ERVEN, P. HARREMOËS. *Rényi Divergence and Kullback-Leibler Divergence*, 2012, Submitted to IEEE Transactions on Information Theory, http://hal.inria.fr/hal-00758191.