IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2013

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, machine learning and statistical methods

Table of contents

Project-Team SELECT

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

Creation of the Project-Team: 2007 January 01.

1. Members

Research Scientists

Gilles Celeux [Inria, Senior Researcher]
Erwan Le Pennec [Ecole Polytechnique, Researcher]

Faculty Members

Pascal Massart [Team leader, Univ. Paris XI, Professor]
Christine Keribin [Univ. Paris XI, Associate Professor]
Patrick Pamphile [Univ. Paris XI, Associate Professor]
Jean-Michel Poggi [Univ. Paris V, Professor, HdR]
Yves Rozenholc [Inria, Associate Professor, from Sep 2013, HdR]

External Collaborators

Yves Auffray [Dassault]
Serge Cohen [Ipanema,CNRS]

Engineers

Benjamin Auder [CNRS]
Yves Misiti [CNRS]

PhD Students

Vincent Brault [Univ. Paris XI]
Émilie Devijver [Univ. Paris XI]
Rémy Fouchereau [Snecma, granted by CIFRE]
Melina Gallopin [Univ. Paris XI]
Clément Levrard [Univ. Paris XI]
Nelo Molter Magalhaes [Univ. Paris XI]
Lucie Montuelle [Univ. Paris XI]
Valérie Robert [Univ. Paris XI, from Oct 2013]
Solenne Thivin [Thales]
Vincent Thouvenot [EDF, granted by CIFRE]
Yann Vasseur [Univ. Paris XI]

Post-Doctoral Fellows

Jairo Cugliari [Inria, granted by EDF SA, until Aug 2013]
Mesrob Ohannessian [Inria, until Sep 2013]
Mohammed Sedki [Inria, granted by Fondation de Cooperation Scientifique Campus Paris Saclay-DIGITEO, until Aug 2013]
Tim Van Erven [Univ. Paris XI]

Administrative Assistant

Katia Evrat [Inria]

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. Research Program

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.

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).

Moreover, a collaboration has started with Dassault Aviation on modal analysis of mechanical structures, which aims at identifying the vibration behavior of structures under dynamic excitations. From algorithmic view point, modal analysis amounts to estimation in parametric models on the basis of measured excitations and structural responses data. As it appears from literature and existing implementations, the model selection problem attached to this estimation is currently treated by a rather heavy and very heuristic proced ure. The model selection via penalisation tools are intended to be tested on this model selection problem.

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

Since many years SELECT collaborates 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. Yann Vasseur started a thesis cosupervised by Gilles Celeux and Marie-Laure Martin-Magniette on this topic which is also an interesting investigation domain for the latent block model developed by SELECT.

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. Two methods for spatial outlier detection have been considered: one based on the nearest neighbours weighted median and one based on kriging increments instead of more traditional pseudo-innovations. The two methods are applied to the PM10 monitoring network in Normandie (France) and are fully implemented in the Measurements Quality Control process.

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 and Platforms

5.1. MIXMOD software

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

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 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 spatially.

Benjamin Auder contributes to the informatics improvement of MIXMOD. He implemented an interface to test any mathematical library (Armadillo, Eigen, ...) to replace NEWMAT. He contributed to the continuous integration setup using Jenkins tool and prepared an automated testing framework for unit and non-regression tests.

5.2. BLOCKCLUSTER software

Participants: Vincent Brault, Gilles Celeux, Christine Keribin.

Mixture model, Block cluster analysis,

Blockcluster is a software devoted on model-based block clustering. It is developed by MODAL team (Inria Lille). With Parmeet Bathia (Inria Lille), Vincent Brault has added a Bayesian point of view for the binary, categorial and continuous datas with the variational Bayes algorithm or Gibbs sampler. Criteria ICL and BIC are used for selecting a relevant block clustering.

6. New Results

6.1. Model selection in Regression and Classification

Participants: Gilles Celeux, Serge Cohen, Jairo Cugliari, Tim Van Erwen, Clément Levrard, Erwan Le Pennec, Pascal Massart, Nelo Molter Magalhaes, Lucie Montuelle, Mohammed Sedki.

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 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.

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.

Lucie Montuelle focused on conditional density estimation by Gaussian mixtures with logistic weights. Using maximum likelihood estimators, a model selection procedure has been applied, supported by a theoretical guarantee. Numerical experiments have been conducted for regression mixtures with parametric logistic weights, using EM and Newton algorithms. This work is available in the research report and a submitted article.

In collaboration with Lucien Birgé (Université Paris 6), Pascal Massart and Nelo Molter Magalhaes define for the algorithm selection problem a new general cross validation procedure based on robust tests, which is an extension of the hold-out defined by Birgé. They get an original procedure based on the Hellinger distance. This procedure is the unique procedure which does not use any contrast function since it does not estimate the risk. They provide theoretical results showing that, under some weak assumptions on the considered statistical methods, the selected estimator satisfies an oracle type inequality. And, they prove that their robust method can be implemented with a sub-quadratic complexity. Simulations show that their estimator performs generally well for estimating a density with different sample sizes and can handle well-known problems, such as histogram or bandwidth selection.

In collaboration with Gérard Biau (Université Paris 6), Clément Levrard and Pascal Massart provide intuitive conditions have been derived for the k-means clustering algorithm to achieve its optimal rate of convergence. They can be thought of as margin conditions such as ones introduced by Mammen and Tsybakov in the statistical learning framework. These conditions can be checked in many cases, such as Gaussian mixtures with a known number of components and do not require the underlying distribution to have a density, on the contrary to the previous fast rates conditions introduced in this domain. Moreover, It allows to derive non-asymptotic bounds on the mean squared distortion of the k-mean estimator, emphasizing the role played by several other parameters of the quantization issue, such as the smallest distance between optimal codepoints or the excess risk of local minimizers. The influence of these parameters is still in discussion, but some previous results show that some of them are crucial for the minimax results obtained in quantization theory.

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 separately. 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. Together with Jairo Cugliari it has also been applied to improve regional forecasts of electrical load consumption using the fact that the consumption of all regions together must add up to the total consumption over the whole country.

The well-documented and consistent variable selection procedure in model-based cluster analysis and classification, that Cathy Maugis (INSA Toulouse) has designed during her PhD. thesis in SELECT, makes use of stepwise algorithms which are painfully slow in high dimensions. In order to circumvent this drawback, Gilles Celeux and Mohammed Sedki, in collaboration with Cathy Maugis, proposed to sort the variables using a lasso-like penalization adapted to the Gaussian mixture model context. Using this rank to select the variables they avoid the combinatory problem of stepwise procedures. Their algorithm is now tested on several challenging simulated and real data sets, showing encouraging performances.

In collaboration with Jean-Michel Marin (Université de Montpellier) and Olivier Gascuel (LIRMM), Gilles Celeux has started a research aiming to select a short list of models rather a single model. This short list of models is declared to be compatible with the data using a p-value derived from the Kullback-Leibler distance between the model and the empirical distribution. And, the Kullback-Leibler distances at hand are estimated trough parametric bootstrap procedures.

6.2. Statistical learning methodology and theory

Participants: Vincent Brault, Gilles Celeux, Christine Keribin, Erwan Le Pennec, Lucie Montuelle, Mesrob Ohannessian, Michel Prenat, Solenne Thivin.

Gilles Celeux, Christine Keribin and the Ph D. student Vincent Brault continued their study on the Latent Block Model (LBM), and worked more especially on categorical data. They further investigated a Gibbs algorithm to avoid solutions with empty clusters on synthetic as well as real data (Congressional Voting Records and genomic data) [STCO13].They detailed the link between the information criteria ICL and BIC, compared them on synthetic and real data, and conjectured that these criteria are both consistent for LBM, which is not a standard behavior. ICL has been proved to be preferred for LBM.

V. Brault applied the Large Gaps algorithm and compared it with other existing algorithms [Aussois13]. He also derived a CEM algorithm for categorical LBM [Agroselect13]. In partnership with the Inria- MODAL team, he implemented the algorithms and information criteria in the R package blockcluster.

C. Keribin has started a collaboration with Tristan Mary-Huard (AgroParisTech) by the supervision of an internship (Master 2) on the use of LBM with truncated Poisson data.

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 approach based on test decision theory. They have obtained theoretical and numerical results on a segmentation based approach in which a simple Markov field testing procedure is used in each cell of a data driven partition.

Erwan Le Pennec and Michel Prenat have also collaborated on a cloud texture modeling using a non-parametric approach. Such a modeling coud be used to better calibrate the detection procedure: it can lead to more examples than the one acquired and it could be the basis of an ensemble method.

Mesrob Ohannessian joined SELECT through an ERCIM Alain Bensoussan fellowship. During his stay, his work focused on two different aspects of statistics: large datasets and data scarcity. In collaboration with researchers in ETH Zurich (Prof. Andreas Krause), he studied the possibility of trading off statistical performance and computational speed in the context of k-means clustering, using the notion of coresets. In collaboration with researchers in Paris 11 (Prof. Elisabeth Gassiat) and Paris 7 (Prof. Stéphane Boucheron), he worked on adaptive universal compression when the alphabet is very large, meaning that some symbol observations are scarce.

6.3. Reliability

Participants: Yves Auffray, Gilles Celeux, Rémy Fouchereau, Patrick Pamphile.

Since 2011, in the framework of a CIFRE convention with Snecma-SAFRAN Rémy Fouchereau has started a thesis on the modeling of fatigue lifetime supervised by Gilles Celeux and Patrick Pamphile. In aircraft, space and nuclear industry, fatigue test is the main basic tool for analyzing fatigue lifetime of a given material, component, or structure. A sample of the material is subjected to cyclic loading S (stress, force, strain, etc.), by a testing machine which counts N, the number of cycles to failure. Fatigue test results are plotted on a SN-curve. A probabilistic model for the construction of SN-curve is proposed. In general, fatigue test results are widely scattered for High Cycle Fatigue region and "duplex" SN-curves appears for Very High Cycle region. That is why classic models from mechanic of rupture theory on one hand, probability theory on the other hand, do not fit SN-curve on the whole range of cycles. We have proposed a probabilistic model, based on a fracture mechanic approach: few parameters are required and they are easily interpreted by mechanic or material engineers.This model has been applied to both simulated and real fatigue test data sets. The SN-curves have been well fitted on the whole range of cycles. The parameters have been estimated using the EM algorithm, combining Newton-Raphson optimisation method and Monte Carlo integral estimations. Recently, the model has been improved taking into account production process information, thanks to a clustering approach. Thus, we have provided engineers with a probabilistic tool for reliability design of mechanical parts, but also with a diagnostic tool for material elaboration.

Since 2013, Gilles Celeux and Patrick Pamphile supervise, in the framework of a collaboration with CEA not yet finalized, a thesis on the modeling of battery State Of Charge for electrical vehicles. Electrical battery is an electrochemical device that converts stored chemical energy into electrical energy. This conversion is reversible and can be repeated during charge/discharge cycles. In an electric vehicle, the battery State Of Charge (SOC) gives the driver indication of how long he can drive without recharging the battery. Unfortunately the complex nature of electrochemical reactions does not allow to measure the SOC directly. Different methods of estimation exist, but they are not robust to various environment conditions (temperature, vehicle driving,...) and to the battery ageing. We propose to estimate the SOC from an *Markov-switching model* : the measurement equation specifies how the SOC depends of an unobservable Markov chain and physical data (temperature, voltage and current intensity,...). Moreover, the SOC estimation is included in the Battery Management System, and therefore estimations must be done online, i.e. with minimum information.

A collaboration has started in 2013 with Dassault Aviation on modal analysis of mechanical structures, which aims at identifying the vibration behavior of structures under dynamic excitations. From algorithmic view point, modal analysis amounts to estimation in parametric models on the basis of measured excitations and structural responses data. As it appears from literature and existing implementations, the model selection problem attached to this estimation is currently treated by a rather heavy and very heuristic procedure. The model selection via penalization tools are intended to be tested on this model selection problem.

6.4. Statistical analysis of genomic data

Participants: Vincent Brault, Gilles Celeux, Christine Keribin.

In collaboration with Florence Jaffrezic and Andrea Rau (INRA, animal genetic department), Mélina Gallopin has started a thesis under the supervision of Gilles Celeux. This thesis is concerned with building statistical networks of genes in animal grenetic. In animal genetic, datasets have a large number of genes and low number of statistical units. For this reason, standard network inference techniques work poorly in this case. At first, this team has developed a data-based method to filter replicated RNA-seq experiments. The method, implemented in the Bioconductor R package `HTSFilter`, removes low expressed genes by optimizing the Jaccard index and reduce the dimension of the dataset. Now, they are studying a clustering model on their expression profiles measured by RNAseq data using Poisson mixture models. External biological knowledge, such as Gene Ontology annotations are taken into account in the model selection step, based on a approximation of the completed log-likelihood given the annotations.

In collaboration with Marie-Laure Martin-Magniette (URGV), GIlles Celeux and Christine Keribin has started a research concerning the buliding statistical networks of transcription factors (TF) with Gaussian Graphical Models (GGM) in the frawork of the intership of Yann Vasseur (Université Paris-sud) who is starting a PhD. thesis on the same subject at the end of 2013. Since the number of TF is greater than the number of statistical units, a lasso-like procedure is used. Moreover the edges of the network are interpreted using the Latent Block Model studied by Vincent Brault in his thesis. An open issue to be solved is the choice of the regularization parameter in the lasso procedure. It is also important to develop this statistical inference for data with good biological control and knowledge to assess the biological relevance of the proposed models.

6.5. Curves classification, denoising and forecasting

Participants: Jairo Cugliari, Émilie Devijver, Pascal Massart, Jean-Michel Poggi, Vincent Thouvenot.

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

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 was strongly related to the use of wavelets together with curves clustering in order to perform accurate load consumption forecasting. The thesis contains methodological and applied aspects linked to the electrical context as well as theoretical ones by introducing external variables in the context of nonparametric forecasting time series. See http://hal.archives-ouvertes.fr/docs/00/78/82/49/PDF/cugliari-jma.pdf and http://hal.inria.fr/docs/00/55/99/39/PDF/RR-7515.pdf The industrial post-doc of Jairo Cugliari, funded by EDF, explores three aspects of this model that complement the original methodology: first, the construction of a confidence interval for the predictor function, second, the flexibility and simplicity of the model to provide, without extra effort, forecasts horizons further and further away and finally, and third: study of the ability to provide good predictions in the presence of subtle signal nonstationarities induced by loss of customers coming from various scenarios, see http://hal.archives-ouvertes.fr/docs/00/81/49/24/PDF/kwf-suite.pdf

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 Émilie 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 her 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. C. Keribin examined the PhD defence of Virgile Fritsch High-dimensional statistical methods for inter-subjects studies in neuroimaging (Inria, Parietal team).

7. Bilateral Contracts and Grants with Industry

7.1. Contract with EDF

Participants: Jairo Cugliari, Jean-Michel Poggi.

SELECT has a contract with EDF regarding wavelet analysis of the electrical load consumption for the aggregation and desagggregation of curves to improve total signal prediction.

7.2. Contract with SNECMA

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.

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 AgroParisTech.

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.

SELECT is animating a working group on Classification, Statistics and fMRI imaging with Neurospin.

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 Bologna (Italy).

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 *Journal de la SFdS*.

  He is Associate Editor of *Statistics and Computing*, *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, Jean-Michel Poggi.

- Gilles Celeux was invited speaker to the annual meeting of the Italian Society of Statistics in Brescia (June 2013), to the joint meeting of the British Classification Society, and the German Classification Society in London (November 2013) and to the Summer Model-Based Clustering working group in Bologna (July 2013).
- Jean-Michel Poggi was invited speaker at , International worshop Unsupervised Learning and High-dimensional Statistics, ENS Ulm, Paris, 11 (September 2013).

9.1.3. Scientific animation

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

- Gilles Celeux was Chair of the Program Committee of the Annual Journées de Statistique de la SFdS, Toulouse, May 2013. Gilles Celeux was organizer of the ERCIM 2013 Session Model Selection, London, 14-16 December 2013.
- 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). Christine Keribin is member of the council of the French statistical society (SFdS). Since this year, she organizes conferences for statisticians from various backgrounds (academic, software, companies) to exchange experiences on statistical methods and software http://www.sfds.asso.fr/323-Rendez_vous_SFdS_Methodes_et_Logiciels.
- 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.
- Jean-Michel Poggi was President of the French statistical society (SFdS).

- Jean-Michel Poggi is member of the EMS (European Mathematical Society) Committee for Applied Mathematics.
- Jean-Michel Poggi is Guest Editor (with R. Kenett, A. Pasanisi) of the special issue on Special Issue on Graphical causality models: Trees, Bayesian Networks and Big Data, in Quality Technology and Quantitative Management (QTQM).
- Jean-Michel Poggi is Editor (with A. Antoniadis, X. Brossat) of a Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension, Springer 2014.
- Jean-Michel Poggi was Member Editorial committee of 2013 Mathématiques pour la planète terre : un jour, une brève, an initiative from Cap'Maths, Inria, INSMI, SFdS, SMAI, SMF for 2013 the year of Mathematics for Planet Earth.
- Jean-Michel Poggi is Organizer and President of the Scientific committee (with R. Kenett, A. Pasanisi) of the ENBIS-SFdS 2014 Spring Meeting on Graphical causality models: Trees, Bayesian Networks and Big Data, IHP, Paris, 9-11April 2013.
- Jean-Michel Poggi is Organizer of the meeting Horizons de la Statistique, Paris, IHP, 21 January 2014.
- Jean-Michel Poggi was organizer of the ERCIM 2013 Session Random forests and related methods: theory and practice, London, 14-16 December 2013

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: 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 in progress: Vincent Brault, 2011, Gilles Celeux and Christine Keribin

PhD in progress: Rémi Fouchereau, 2011, Gilles Celeux and Patrick Pamphile

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: 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

9.3. Popularization

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

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] E. LE PENNEC. , *Some (statistical) applications of Ockham's principle*, Université Paris Sud - Paris XI, March 2013, Habilitation à Diriger des Recherches, http://hal.inria.fr/tel-00802653

Articles in International Peer-Reviewed Journals

[2] S. X. COHEN, E. LE PENNEC. *Partition-Based Conditional Density Estimation*, in "ESAIM: Probability and Statistics", 2013, vol. 17, pp. 672–697 [*DOI :* 10.1051/PS/2012017], http://hal.inria.fr/hal-00915854

[3] M. GALLOPIN, A. RAU, F. JAFFRÉZIC. *A Hierarchical Poisson Log-Normal Model for Network Inference from RNA Sequencing Data*, in "PLoS ONE", October 2013, http://hal.inria.fr/hal-00921397

[4] A. RAU, M. GALLOPIN, G. CELEUX, F. JAFFREZIC. *Data-based filtering for replicated high-throughput transcriptome sequencing experiments*, in "Bioinformatics", 2013, vol. 29, pp. 2146-2152 [*DOI :* 10.1093/BIOINFORMATICS/BTT350], http://hal.inria.fr/hal-00927025

[5] 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", 2013, vol. 64, pp. 220-236, http://hal.inria.fr/inria-00516991

Invited Conferences

[6] G. CELEUX. *How useful Bayesian inference could be in Model-based clustering?*, in "Advances in Latent Variables-Methods, Models and Applications", Brescia, Italy, Società italiana di statistica, June 2013, http://hal.inria.fr/hal-00927006

[7] G. CELEUX. *The notion of alpha-admissible models: illustrations in the model-based clustering context*, in "Working Group on Model-Based Clustering", Bologne, Italy, University of Bologna, July 2013, http://hal.inria.fr/hal-00926997

[8] G. CELEUX. *Variable selection in clustering and classification: issues, difficulties and solutions*, in "AG DANK/BCS Meeting 2013", London, United Kingdom, C. HENNIG (editor), German and British Classification Societies, November 2013, http://hal.inria.fr/hal-00927011

International Conferences with Proceedings

[9] L. MONTUELLE, E. LE PENNEC. *R égression gaussienne à poids logistiques et maximum de vraisemblance p énalis é*, in "JDS-45e Journées de Statistique", Toulouse, France, May 2013, http://hal.inria.fr/hal-00921524

Conferences without Proceedings

[10] V. BRAULT. *Modèle des blocs latents et algorithme Largest Gaps*, in "Cinquièmes Rencontres des Jeunes Statisticien-ne-s", Aussois, France, Jérôme Saracco, August 2013, http://hal.inria.fr/hal-00924395

[11] V. BRAULT, G. CELEUX, C. KERIBIN. *Comparaisons de différents algorithmes pour le modèle des blocs latents*, in "Séminaire AgroSelect", Paris, France, October 2013, http://hal.inria.fr/hal-00924404

[12] M. GALLOPIN. *Gene network inference from miRNA-seq data*, in "MicroRNA Workshop "MicroRNA Technology, Relevance and Application"", Bruxelles, Belgium, May 2013, http://hal.inria.fr/hal-00921413

[13] M. GALLOPIN. *Inferring gene networks from RNA-seq data*, in "5th workshop statseq", Helsinki, Finland, April 2013, http://hal.inria.fr/hal-00921449

[14] C. KERIBIN, V. BRAULT. *Model selection with untractable likelihood*, in "ERCIM - 6th International Conference of the ERCIM Working Group on Computing and Statistics, 2013", London, United Kingdom, December 2013, http://hal.inria.fr/hal-00924197

Research Reports

[15] C. KERIBIN, V. BRAULT, G. CELEUX, G. GOVAERT. , *Estimation and Selection for the Latent Block Model on Categorical Data*, Inria, March 2013, n^{o} RR-8264, 31 p. , http://hal.inria.fr/hal-00802764

[16] M. MISITI, Y. MISITI, J.-M. POGGI, B. PORTIER. , *Mixture of linear regression models for short term PM10 forecasting in Haute Normandie (France)*, February 2013, http://hal.inria.fr/hal-00841349

[17] L. MONTUELLE, E. LE PENNEC, S. COHEN. , *Gaussian Mixture Regression model with logistic weights, a penalized maximum likelihood approach*, Inria, April 2013, n^{o} RR-8281, http://hal.inria.fr/hal-00809735

Other Publications

[18] A. ANTONIADIS, X. BROSAT, J. CUGLIARI, J.-M. POGGI. , *Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité*, 2013, 1 p. , http://hal.inria.fr/hal-00814530

[19] G. CELEUX, M.-L. MARTIN-MAGNIETTE, C. MAUGIS-RABUSSEAU, A. E. RAFTERY. , *Comparing Model Selection and Regularization Approaches to Variable Selection in Model-Based Clustering*, 2013, http://hal.inria.fr/hal-00849439

[20] J. CUGLIARI. , *Conditional Autoregressive Hilbertian processes*, 2013, http://hal.inria.fr/hal-00788249

[21] R. FOUCHEREAU, G. CELEUX, P. PAMPHILE. , *Probabilistic modeling of S-N curves*, January 2014, http://hal.inria.fr/hal-00924080

[22] C. LEVRARD. , *Margin conditions for vector quantization*, 2013, 42 p. , http://hal.inria.fr/hal-00877093

[23] C. LEVRARD. , *Non Asymptotic Bounds for Vector Quantization*, 2013, 27 p. , Technical proofs are omitted and can be found in the related unpublished paper "Margin conditions for vector quantization", http://hal.inria.fr/hal-00877564

[24] L. RÉMI, I. SERGE, L. FLORENT, C. BIERNACKI, G. CELEUX, G. GOVAERT. , *Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library*, 2013, http://hal.inria.fr/hal-00919486

[25] A. SAUMARD. , *Optimal model selection in heteroscedastic regression using piecewise polynomials*, 2013, http://hal.inria.fr/hal-00512306

[26] T. VAN ERVEN, J. CUGLIARI. , *Game-theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts*, December 2013, http://hal.inria.fr/hal-00920559

[27] S. DE ROOIJ, T. VAN ERVEN, P. GRÜNWALD, W. KOOLEN. , *Follow the Leader If You Can, Hedge If You Must*, December 2013, http://hal.inria.fr/hal-00920549