IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2014

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, Professor]
Yves Rozenholc [Univ. Paris V, Associate Professor, HdR]

Faculty Members

Pascal Massart [Team leader, Univ. Paris XI, Professor]
Christine Keribin [Univ. Paris XI, Associate Professor]
Claire Lacour [Univ. Paris XI, Associate Professor]
Patrick Pamphile [Univ. Paris XI, Associate Professor]
Jean-Michel Poggi [Univ. Paris V, Professor]

Engineers

Benjamin Auder [CNRS, Engineer]
Yves Misiti [CNRS]

PhD Students

Clément Levrard [Univ. Paris XI, until Sep 2014]
Vincent Brault [Univ. Paris XI, until Sep 2014]
Émilie Devijver [Univ. Paris XI]
Rémy Fouchereau [Snecma,granted by Cifre, until Mar 2014]
Mélina Gallopin [Univ. Paris XI]
Jana Kalawoun [CEA]
Nelo Molter Magalhaes [Univ. Paris XI, until Sep 2014]
Lucie Montuelle [Univ. Paris XI, until Dec 2014]
Valérie Robert [Univ. Paris XI]
Solenne Thivin [Thales, granted by Cifre]
Vincent Thouvenot [EDF, granted by Cifre]
Yann Vasseur [Univ. Paris XI]

Post-Doctoral Fellow

Tim Van Erven [Univ. Paris XI, until Mar 2014]

Administrative Assistants

Céline Halter [Univ. Strasbourg]
Olga Mwana Mobulakani [Inria]

Others

Yves Auffray [Dassault]
Serge Cohen [Ipanema,CNRS]
Michel Prenat [Thales]

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 medical images) 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 image analysis.

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. Dynamic contrast Enhanced imaging

Since Yves Rozenholc joins SELECT, we are involved in quantifying tumor microcirculation to monitor treatments in cancer. Dynamic Contrast Enhanced (DCE) imaging provides information on the qualities of a vascular network. It enables biostatisticians to design biomarkers that can be used for diagnosis, prognosis and treatment monitoring. To make available robust tumoral microcirculation biomarkers in DCE imaging, Yves Rozenholc is developing several tools for denoising and clustering the dynamics found in DCE imaging sequences, to realize in the blood flow model, and testing equality of the survival functions coming from two DCE imaging sequences.

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. On an other hand, SELECT is involved in ANR "jeunes chercheurs" MixStatSeq directed by Cathy Maugis (INSA Toulouse) wich is concerned with Statistical analysis and clustering of RNASeq genomics data.

4.6. Pharmacovigilance

A collaboration has started with Pascale Tubert-Bitter, Ismael Ahmed and Mohamed Sedki (Pharmacoepidemiology and Infectious Diseases, PhEMI) for the analysis of pharmacovigilance data. In this framework, the objective is to detect as soon as possible potential associations between some drugs and adverse effects which appeared after the authorisation marketing of these drugs. Instead of working on aggregated data (contingency table) like it is usually the case, the developed approach aims at dealing with the individual data which perhaps give more information. Valerie Robert started a thesis cosupervised by Gilles Celeux and Christine Keribin on this topic which enables to develop a new model based-clusering inspired of the latent block model.

4.7. Environment

A study has been achieved by Jean-Michel Poggi, Benjamin Auder and Bruno Portier (INSA de Rouen), in the context of a collaboration between AirNormand, Orsay University and INSA of Rouen. It is an application of sequential prediction. To build the prediction, the question is to optimally combine before every term of forecast, the predictions of a set of experts. The study is original not only because of the specific field of application and the adaptation to the concrete context of the work of the air quality monitor in regional agency, but the main originality is that the initial set of experts contains at the same time experts coming from statistical models built by means of different methods and of different predictors and from experts coming from deterministic physico-chemical models. The interest of this kind of sequential prediction method in this specific context is under investigation and the first results on three monitoring stations are promising.

4.8. 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. New 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 2010, MIXMOD has a proper graphical user interface. 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.

This year, it has been decide to create MIXMODSTORE which proposes companion programs of MIXMOD. As a matter of fact, the program MixmodCombi of Jean-Patrick Baudry (Université Paris 6) and Gilles Celeux which allows a hierarchical clustering derived from a mixture has been associated to Rmixmod.

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. It ha been enriched by a full Bayesian version using a Gibbs sampler. This Gibbs sampler coulpled with the variaotional Bayes algorithm provides solutions more stable and less dependent of the starting values of the algorithm. An exact expression of criterion ICL has been provided. This criterion or 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, Clément Levrard, Erwan Le Pennec, Pascal Massart, Nelo Molter Magalhaes, Lucie Montuelle.

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. With Serge Cohen and an X intern some progress have been made on the use of logistic weights in the hyperspectral setting.

Lucie Montuelle has studied a model of mixture of Gaussian regressions in which the proportions are modeled using 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 published in Electronic Journal of Statistics.

Another subject considered by Erwan Le Pennec and Lucie Montuelle was the obtention of oracle inequalities in deviation for model selection aggregation in the fixed design regression framework. Exponential weights are widely used but sub-optimal. They aggregate linear estimators and penalize Stein's unbiased risk estimate used in exponential weights to derive such inequalities. Furthermore if the infinity norm of the regression function is known and taken into account in the penalty, then a sharp oracle inequality is available. Pac-Bayesian tools and concentration inequalities play a key role in this work. These results may be found in a prepublication on arxiv or in Lucie Montuelle's PhD thesis.

In collaboration with Sylvain Arlot, Matthieu Lerasle an Patricia Reynaud-Bourret (CNRS) Nelo Molter Magalhaes considers estimator selection problem with the L^2 loss. They provide a theoretical minimal and optimal penalty. They define practical cross-validation procedures and provide non-asymptotic and first order optimal results for these procedures.

Emilie Devijver and Pascal Massart focused on the Lasso for high dimension finite mixture regression models. An ℓ_1 oracle inequality have been get for this estimator for this model, for a specific regularization parameter. Moreover, for maximum likelihood estimators, restricted to relevant variables and to low rank, theoretical results have been proved to support methodology.

Pascal Massart and Clément Levrard continue their work on the properties of the k-means algorithm in collaboration with Gérard Biau (Université Paris 6). Most of the work achieved this year was devoted to the obtention of fast convergence rates for the k-means quantizer of a source distribution in the high-dimensional case. It has been proved that the margin condition for vector quantization introduced last year can be extended to the infinite dimensional Hilbert case, and that this condition is sufficient for the source distribution to satisfy some natural properties, such as the finiteness of the set of optimal quantizers. When this condition is satisfied, a dimension-free fast convergence rate can be derived. In addition, this margin condition provides theoretical guarantees for methods combining k-means and variable selection through a Lasso-type procedure. Its implementation is still in process, however early experiments shows that this procedure can retrieve active variables in the Gaussian mixture case.

Among selection methods for nonparametric estimators, a recent one is the procedure of Goldenshluger-Lespki. This method proposes a data-driven choice of m to select an estimator among a collection $(\widehat{s}_m)_{m\in M}$. The selected $\widehat{m}$ is chosen as a minimiser of $B(m)+V(m)$ where $B(m)=\sup\{[\|\widehat{s}_m-\widehat{s}_{m'}\|-V(m')]_+,\ m'\in M\}$ and $V(m)$ is a penalty term to be suitably chosen. Previous results have established oracle inequalities to ensure that if $V(m)$ is large enough the final estimator $\widehat{s}_{\widehat{m}}$ is almost as efficient as the best one in the collection. The aim of the work of Claire Lacour and Pascal Massart was to give a practical way to calibrate $V(m)$. To do this they have evidenced an explosion phenomenon: if V is chosen smaller than some critical V_0, the risk $\|s-\widehat{s}_{\widehat{m}}\|$ is proven to dramatically increase, though for $V>V_0$ this risk is quasi-optimal. Simulations have corroborated this behavior.

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 in collaboration with Mohammed Sedki (Université Paris XI) and 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. After tests on challenging simulated and real data sets, their algorithm finalised and show good performances.

In collaboration with Jean-Michel Marin (Université de Montpellier) and Olivier Gascuel (LIRMM), Gilles Celeux has continued 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 through non parametric and parametric bootstrap procedures.

6.2. Statistical learning methodology and theory

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

Vincent Brault, Ph D. student of Gilles Celeux and Christine Keribin defended his thesis on the Latent Block Model (LBM) for categorical data. Their work investigated a Gibbs algorithm to avoid solutions with empty clusters on synthetic as well as real data (Congressional Voting Records and genomic data. 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. Hence, ICL has to be preferred for LBM. This work is now published in *Statistics and Computing*.

Vincent Brault has achieved a detailed bibliographical review on coclustering with Aurore Lomet (UTC) which is currently under revision. He has also worked in collaboration with Mahindra Mariadassou (INRA) to overview the state of the art on theoretical results for latent or stochastic block model.

Vincent Brault, Christine Keribin and Mahindra Mariadassou have started a collaboration to tackle the consistency and asymptotic normality for the maximum likelihood and variational estimators in a stochastic or latent block model.

Gilles Celeux has started a collaboration with Jean-Patrick Baudry on strategies to avoid the traps of the EM algorithm in mixture analysis. They anayse the effect of the spurious local maximizers and the regulariszed algorithms to avoid these spurious solutions. They explore the link of the degree of regularization and the slope heuristics. Moreover, they propose and study strategies to initiate the EM algorithm embedding the solution with K components and the starting position with $K + 1$ component to avoid suboptimal solutions.

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. They also have obtained experimental results on images (or patches) unsupervised classification, with the aim of better calibrate the detection procedure. The classification is based on features which are defined in cloud texture modeling activity.

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.

6.3. Reliability

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

In 2014, in the framework of a CIFRE convention with Snecma-SAFRAN Rémy Fouchereau has defended 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 SNcurve. 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. The model has been then 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 two years SELECT collaborates with CEA for the estimation of the battery State of Charge (SoC). For vehicles powered by an electric motor, SoC estimation is essential to guarantee vehicle autonomy, as well as safe utilization. The aim is to create a reliable SoC model to closely fit the battery dynamic, in embedded applications (e.g. Electric Vehicle). Jana Kalawoun started a thesis supervised by Gilles Celeux, Patrick Pamphile and Maxime Montaru (CEA) on this topic. The SoC is modeled by a Switching Markov State-Space Model. The parameters are estimated by combining the EM algorithm and Particle Filter methods. The model is validated using real-life electric vehicle data. It has been proved to be highly superior to a simple state space model. The optimal number of battery modes is then identified, using different model selection criteria as BIC or the slope heuristics.

Yves Auffray and Gilles Celeux proposed a solution to a reliability problem on Dassault's F7X business jet brakes. As the origin brake version showed poor reliability performance, an increased frequency inspection of the brakes had been decided and, after a while, a new brake version adopted. The new version has not shown any failure since its adoption. Then the question was : is it possible to relax the brakes inspection frequency ? On the basis of first brake version failure data, the parameters of a Weibull law was estimated : $\eta = 3169, \beta = 1.38$. Under the hypothesis that the new brake version would follow the same Weibull law, the probability that none of them broke was $1.67\ 10^{-6}$. This led to reject that hypothesis.
A Weibull model for the new brakes was then estimated. The shape parameter beeing leaved conservatively unchanged, the scale parameter was estimated so that the no failure event probability amounts to 0.05. This led to $\eta = 9326$.
From the resulting Weibull model, dates $D_0, D_1, \cdots, D_k, \cdots$ of inspection for the new brakes was established so that : $\mathbb{P}(T \leq D_0 + D_1 + \cdots + D_k | T > D_0 + \cdots + D_{k-1}) = 0.01$.
Dassault has adopted this far less constraining inspection calendar.

6.4. Statistical analysis of genomic data

Participants: Vincent Brault, Gilles Celeux, Mélina Gallopin, Christine Keribin, Yann Vasseur.

In collaboration with Florence Jaffrezic and Andrea Rau (INRA, animal genetic department), Mélina Gallopin is a third year PhD student under the supervision of Gilles Celeux. This thesis is concerned with the modelization and model selection in the analysis of RNA-seq data. This year, they proposed a model selection criterion for model-based clustering of annotated gene expression data. This criterion is a ICL-like criterion taking into account the annotations. They are also working on a objective comparison of discrete and continuous modelling after a transformations for RNA-seq data based on a comparison of the likelihoods (eventually penalized) of the models in competition.

The subject of Yann Vasseur PhD Thesis, supervised by Gilles Celeux and Marie-Laure Martin-Magniette (INRA URGV), is the inference of a regulatory network on Transcriptions Factors (TFs), which are specific genes, of *Arabidopsis thaliana*. In that purpose, a transciptome dataset with a sensibly equal size of TFS and statistical units is available. The first aim consists of reducing the dimension of the network to avoid high dimension difficulties. Representing this network with a Gaussian Graphical Model, the following procedure has been defined:

1. *Selection step*: choosing the set of TFs regulators (supports) of each TF.
2. *Classification step*: deducing co-factors groups (TFs with similary expression levels) from these supports.

Thus, the reduced network would be built on the co-factors groups. Currently, several selection methods based on Gauss-LASSO and resampling procedures have been applied on the dataset. The study of the stability and the parameters calibration of these methods are in progress. The TFs are clustered with the Latent Block Model in a number of co-factors groups selected with the BIC or the exact ICL criterion.

In collaboration with Marie-Laure Martin-Magniette, Cathy Maugis and Andrea Rau, Gilles Celeux studied gene expression gotten from high-throughput sequencing technology. They focus on the question of clustering digital gene expression profiles as a means to discover groups of co-expressed genes. They propose a Poisson mixture model using a rigorous framework for parameter estimation as well as the choice of the appropriate number of clusters. They illustrate co-expression analyses using this approach on two real RNA-seq datasets. A set of simulation studies also compares the performance of the proposed model with that of several related approaches developed to cluster RNA-seq or serial analysis of gene expression data. The proposed method is implemented in the open-source R package `HTSCluster`, available on CRAN.

6.5. Model based-clustering for pharmacovigilance data

Participants: Gilles Celeux, Christine Keribin, Valérie Robert.

In collaboration with Pascale Tubert-Bitter, Ismael Ahmed and Mohamed Sedki, Gilles Celeux and Christine Keribin has started a research concerning the detection of associations between drugs and adverse events in the framework of the PhD of Valerie Robert. At first, this team has developed a model-based clustering inspired of the latent black model which consists in co-clustering rows and columns of two binary tables imposing the same row ranking. Then it enables to highlight subgroups of individuals sharing the same drug profile and subgroups of adverse effects and drugs with strong interaction. Besides, some sufficient conditions are provided to obtain the identifiability of the model and some studies are experimented on simulated data.

6.6. Curves classification, denoising and forecasting

Participants: 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, 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, which is dedicated to curves clustering for the prediction. A natural framework to explore this question is mixture of regression models for functional data. They extend to functional data the recent work by Bühlmann and coauthors dealing with the simultaneous estimation of mixture regression models in the scalar case using Lasso type methods. It is 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. Nevertheless, they also propose a procedure dealing with low rank estimator. Simulations and benchmark data have been conducted for high-dimensional finite mixture regression models.

Jean-Michel Poggi, co-supervising with Meriem Jaëdane, Raja Ghozi (ENIT Tunisie) and from the industrial side, Sylvie Sevestre-Ghalila (CEA LinkLab), the PhD thesis of Neska El Haouij, funded by a kind of CIFRE with CEA LinkLab. The industrial motivation of this work is the recent development of new technologies for sensory measurements, environmental and physiological to explain and improve the driving tasks . The thesis aims to explain sensory aspects involved in automated decision to the car interior, by objectivization. The thesis will focus on the use and development of experimental designs and statistical methods to quantify and explain driving ability in to the modeling using functional explanatory factors. Statistical contributions of this work will involve nonparametric estimation and variable selection and/or models.

6.7. Statistical analysis of medical images

Participants: Christine Keribin, Yves Rozenholc.

Yves Rozenholc and C. Keribin work the genomic tumoral alterations and supervised a Master student Yi LIU. The study of genomic DNA alterations (recurrent regions of alteration, patterns of instability) contributes to tumor classification, and becomes of great importance for the personalization of cancer treatments. The use of Single-Nucleotide Polymorphism (SNP) arrays or of New Generation Sequences (NGS) techniques allows the simultaneous estimation of segmented copy number (CN) and B-allele frequency (BAF) profiles along the whole genome. In this context, Popova (2009) proposed the GAP method, based on pattern recognition with (BAF, CN) maps to detect genotype status of each segment in complex tumoral genome profiles. It takes into account the fact that the observations on these maps are necessarily placed on centers that depend –up to a proper scaling of the CN– only on the unknown proportion of non tumoral tissue in the sample. Being deterministic and manually tuned, this method appears sensitive to noise. To overcome this drawback, they set a mixture model, allowing the automatic estimation of the proportion of non tumoral tissue and the test of genotype for each segment along the whole genome. They develop the estimation with an adapted EM algorithm that has been tested on simulated data. This work has already been presented (ERCIM 14, SEQBIO14) and provides many potential developments.

7. Bilateral Contracts and Grants with Industry

7.1. Contract with SNECMA

Participants: Gilles Celeux, Rémy Fouchereau, Patrick Pamphile.

SELECT has a contract with SAFRAN - SNECMA, an high-technology group (Aerospace propul- sion, Aicraft equipment, Defense Security, Communications),regarding modelling reliability of Air- craft Equipment.

7.2. Contract with Thales

Participants: Erwan Le Pennec, Michel Prenat, Solenne Thivin.

SELECT has a contract with Thales Optronique on target detection on complex backgrounds.

8. Partnerships and Cooperations

8.1. Regional Initiatives

Pascal Massart is co-organizing a working group at ENS (Ulm) on Statistical Learning.

Christine Keribin is animating the bimensual rendez-vous SFdS "methods and Software".

Gilles Celeux and Christine Keribin has started a collaboration with the Pharmacoepidemiology and Infectious Diseases (PhEMI, INSERM).

8.2. National Initiatives

8.2.1. ANR

SELECT is participating to the ANR MixStatSeq.

8.3. International Initiatives

Gilles Celeux is one of the co-organizers of the international Working Group on Model-Based Clustering. This year this workshop took place in Dublin (Ireland).

Yves Rozenholc has been invited at the Department of Statistics of the University of Haifa for three weeks, at the Department of Mathematics of Eindhoven University for one week and at the Institut of statistic, biostatistic and actuarial sciences of the catholic University of Louvain.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

Gilles Celeux is one of the co-organizers of the international Working Group on Model-Based Clustering. This year this workshop took place in Dublin (Ireland).

Jean-Michel Poggi was 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 was Editor (with A. Antoniadis, X. Brossat) of a Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension, Springer.

Jean-Michel Poggi was 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, 9-11 April 2014.

Jean-Michel Poggi was Organizer of the meeting Horizons de la Statistique, Paris, IHP, 21 January 2014.

Jean-Michel Poggi was organizer of the ERCIM 2014 Session ElectricityLoad Forecsting, Pisa, 6-8 December 2014.

Yves Rozenholc was the scientific coordinator and organizer of the third edition of the school “Tumoral Genome Analysis”, 12-19 Mai 2014.

9.1.2. Journal

9.1.2.1. Member of the editorial board

Gilles Celeux is Editor-in-Chief of *Journal de la SFdS*. He is Associate Editor of *Statistics and Computing*, *CSBIGS*.

Pascal Massart is Associate Editor of *Annals of Statistics*, *Confluentes Mathematici*, and *Foundations and Trends in Machine Learning*.

Jean-Michel Poggi is Associate Editor of *Journal of Statistical Software*, *Journal de la SFdS* and *CSBIGS*.

9.1.2.2. Reviewer

The members of the team reviewed numerous papers for numerous international journals.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

All the SELECT members are teaching in various courses of different universities and in particular in the Master 2 “Modélisation stochastique et statistique” of University Paris-Sud.

9.2.2. Supervision

PhD: Vincent Brault, Estimation et sélection de modèle pour le modèle des blocs latents, Université Paris-Sud, September 2014, Gilles Celeux and Christine Keribin

PhD: Rémi Fouchereau, Modélisation probabiliste des courbes S-N, Université Paris-Sud, March 2014, Gilles Celeux and Patrick Pamphile

PhD: Lucie Montuelle, Inégalités d’oracle et mélanges, Université Paris-Sud, December 2014, Erwan Le Pennec

PhD: Clément Levrard, Quantification vectorielle en grande dimension : vitesses de convergence et sélection de variables, Université Paris-Sud, September 2014, Pascal Massart and Gérard Biau (UPMC)

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

PhD in progress: Jana Kalawoun, 2012, Gilles Celeux et Patrick Pamphile

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

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

PhD in progress: Valérie Robert, 2013, Gilles Celeux et Christine Keribin

PhD in progress: Yann Vasseur, 2013, Gilles Celeux et Marie-Laure Martin-Magniette (URGV)

PhD in progress: Neska El Haouij, 2014, Jean-Michel Poggi and Meriem Jaïdane, Raja Ghozi (ENIT Tunisie) and Sylvie Sevestre-Ghalila (CEA LinkLab), Thesis ENITUPS

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

9.3. Popularization

Gilles Celeux and Valérie Robert have written an article on statistics in basket-ball to appear in the Journal of the SFdS, special issue ’sport and statistics’.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] V. BRAULT. *Estimation and model selection for the latent block model*, Université Paris sud 11, September 2014, https://hal.inria.fr/tel-01090340

[2] R. FOUCHEREAU. *Probabilistic modelling of S-N curves*, Université Paris Sud - Paris XI, April 2014, https://tel.archives-ouvertes.fr/tel-00990770

[3] C. LEVRARD. *High-dimensional vector quantization: convergence rates and variable selection*, Universite de Paris 11, September 2014, https://tel.archives-ouvertes.fr/tel-01093476

[4] L. MONTUELLE. *Oracle inequalities and mixtures*, Université Paris-Sud, December 2014, https://hal.inria.fr/tel-01109103

Articles in International Peer-Reviewed Journals

[5] Y. AUFFRAY, P. BARBILLON, J.-M. MARIN. *Bounding rare event probabilities in computer experiments*, in "Computational Statistics and Data Analysis", July 2014, 24 p. [*DOI :* 10.1016/J.CSDA.2014.06.023], https://hal.inria.fr/hal-01097166

[6] J.-P. BAUDRY, M. CARDOSO, G. CELEUX, M.-J. AMORIM, A. SOUSA FERREIRA. *Enhancing the selection of a model-based clustering with external categorical variables" Advances in Data Analysis and Classification*, in "Advances in Data Analysis and Classification", 2014, 14 p. , https://hal.inria.fr/hal-01108795

[7] G. CELEUX, M.-L. MARTIN-MAGNIETTE, C. MAUGIS, A. E. RAFTERY. *Comparing Model Selection and Regularization Approaches to Variable Selection in Model-Based Clustering*, in "Journal de la Société Française de Statistique", 2014, https://hal.inria.fr/hal-00943473

[8] R. FOUCHEREAU, G. CELEUX, P. PAMPHILE. *Probabilistic modeling of S-N curves*, in "International Journal of Fatigue", 2014, 10 p. , https://hal.inria.fr/hal-01108802

[9] S. FU, G. CELEUX, N. BOUSQUET. *Bayesian inference for inverse problems occurring in uncertainty analysis*, in "International Journal for Uncertainty Quantification", 2014, 12 p. , https://hal.inria.fr/hal-01108811

[10] C. KERIBIN, V. BRAULT, G. CELEUX, G. GOVAERT. *Estimation and selection for the latent block model on categorical data*, in "Statistics and Computing", 2014, 16 p. [*DOI :* 10.1007/s11222-014-9472-2], https://hal.inria.fr/hal-01095957

[11] R. LEBRET, S. IOVLEFF, F. LANGROGNET, C. BIERNACKI, G. CELEUX, G. GOVAERT. *Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library*, in "Journal of Statistical Software", December 2014, forthcoming, https://hal.archives-ouvertes.fr/hal-00919486

[12] L. MONTUELLE, E. LE PENNEC. *Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach*, in "Electronic Journal of Statistics", September 2014, vol. 8, n^{o} 1, 35 p. [*DOI :* 10.1214/14-EJS939], https://hal.inria.fr/hal-01101483

[13] A. RAU, C. MAUGIS-RABUSSEAU, M.-L. MARTIN-MAGNIETTE, G. CELEUX. *Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models*, in "Bioinformatics", 2015, 15 p. [*DOI :* 10.1093/BIOINFORMATICS/BTU845.], https://hal.inria.fr/hal-01108821

[14] T. VAN ERVEN, J. CUGLIARI. *Making Regional Forecast add up*, in "Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension", 2014, https://hal.inria.fr/hal-00943529

Articles in National Peer-Reviewed Journals

[15] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité*, in "Journal de la Société Française de Statistique", 2014, vol. 155, n^{o} 2, pp. 202- 219, ISSN: 2102-6238 - Open access, https://hal.inria.fr/hal-00942680

International Conferences with Proceedings

[16] M. GALLOPIN, G. CELEUX, F. JAFFRÉZIC, A. RAU. *A model selection criterion for unsupervised model-based clustering of annotated data: applications to the analysis of RNA-seq* , in "XXVII International Biometric Conference", Florence, Italy, July 2014, https://hal.inria.fr/hal-01096332

[17] R. GENUER, J.-M. POGGI, C. TULEAU-MALOT. *VSURF : un package R pour la sélection de variables à l'aide de forêts aléatoires*, in "46èmes Journées de Statistique", Rennes, France, 2014, https://hal.inria.fr/hal-01096233

[18] C. KERIBIN, Y. LIU, T. POPOVA, Y. ROZENHOLC. *Statistical quantification of genomic tumoral alterations with a mixture model*, in "ERCIM 2014", Pisa, Italy, December 2014, https://hal.inria.fr/hal-01095984

[19] L. MONTUELLE, E. LE PENNEC. *Agrégation PAC-bayésienne d'estimateurs par projection*, in "46e Journées de Statistique", Rennes, France, SFdS, June 2014, https://hal.inria.fr/hal-01097173

National Conferences with Proceedings

[20] V. BRAULT, G. CELEUX, C. KERIBIN. *Implementation of the Gibbs sampler for the latent block model*, in "46èmes journées de statistique de la SFdS", Rennes, France, SFdS, June 2014, https://hal.inria.fr/hal-01090349

[21] M. GALLOPIN, G. CELEUX, F. JAFFRÉZIC, A. RAU. *Un critère de sélection de modèle pour la classification non supervisée de données annotées: applications à l'analyse de données d'expression de gènes RNA-seq*, in "46èmes Journées de Statistique de la SFdS", Rennes, France, June 2014, https://hal.inria.fr/hal-01096346

Conferences without Proceedings

[22] R. GENUER, J.-M. POGGI, C. TULEAU-MALOT. *VSURF : un package R pour la sélection de variables à l'aide de forêts aléatoires*, in "3èmes Rencontres R", Montpellier, France, 2014, https://hal.inria.fr/hal-01096237

[23] A. TORRADO-CARVAJAL, A. HERNANDEZ-TAMAMES J., L. HERRAIZ J., Y. ERYAMAN, Y. ROZENHOLC, E. ADALSTEINSSON, L. WALD, N. MALPICA. *A Multi-Atlas and Label Fusion Approach for Patient-Specific MRI Based Skull Segmentation*, in "International Society for Magnetic Resonance in Medicine (ISMRM)", Milan, Italy, May 2014, https://hal.archives-ouvertes.fr/hal-01108453

Scientific Books (or Scientific Book chapters)

[24] P. MASSART. *A non asymptotic walk in probability and statistics*, in "Past, Present, and Future of Statistical Science", X. LIN, C. GENEST, D. L. BANKS, G. MOLENBERGHS, D. W. SCOTT, J.-L. WANG (editors), Chapman and Hall/CRC, 2014, https://hal.inria.fr/hal-00942827

Research Reports

[25] G. CELEUX, V. ROBERT. *Towards an objective team efficiency rate in basketball*, July 2014, 19 p. , https://hal.inria.fr/hal-01020295

[26] M. SEDKI, G. CELEUX, C. MAUGIS. *SelvarMix: A R package for variable selection in model-based clustering and discriminant analysis with a regularization approach*, August 2014, 17 p. , https://hal.inria.fr/hal-01053784

Other Publications

[27] A. ANTONIADIS, X. BROSSAT, J. CUGLIARI, J.-M. POGGI. *A prediction interval for a function-valued forecast model*, August 2014, https://hal.archives-ouvertes.fr/hal-01094797

[28] A. BAR HEN, S. GEY, J.-M. POGGI. *Influence functions for CART*, February 2014, Preprint HAL, https://hal.inria.fr/hal-00944098

[29] E. DEVIJVER. *An ℓ 1 -oracle inequality for the Lasso in finite mixture of multivariate Gaussian regression models*, October 2014, https://hal.inria.fr/hal-01075338

[30] E. DEVIJVER. *Finite mixture regression: a sparse variable selection by model selection for clustering*, September 2014, 20 pages, https://hal.archives-ouvertes.fr/hal-01060079

[31] E. DEVIJVER. *Model-based clustering for high-dimension data. Application to functional data*, September 2014, 20 pages, https://hal.archives-ouvertes.fr/hal-01060063

[32] E. DEVIJVER. *Joint rank and variable selection for parsimonious estimation in high-dimension finite mixture regression model*, January 2015, https://hal.archives-ouvertes.fr/hal-01099296

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

[34] P. GAILLARD, G. STOLTZ, T. VAN ERVEN. *A Second-order Bound with Excess Losses*, February 2014, https://hal.archives-ouvertes.fr/hal-00943665

[35] M. GALLOPIN, G. CELEUX, F. JAFFRÉZIC, A. RAU. *A model selection criterion for model-based clustering of annotated gene expression data*, November 2014, https://hal.inria.fr/hal-01088870

[36] C. LEVRARD. *Margin conditions for vector quantization*, April 2014, 43 pages, https://hal.archives-ouvertes.fr/hal-00877093

[37] C. LEVRARD. *Variable selection for k means quantization*, June 2014, https://hal.archives-ouvertes.fr/hal-01005545

[38] N. MAGALHÃES, Y. ROZENHOLC. *An efficient algorithm for T-estimation*, April 2014, https://hal.archives-ouvertes.fr/hal-00986229

[39] L. MONTUELLE, E. LE PENNEC. *PAC-Bayesian aggregation of linear estimators*, September 2014, https://hal.inria.fr/hal-01070805