IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2015

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

Creation of the Project-Team: 2007 January 01

Keywords:

Computer Science and Digital Science:

- 3.1.1. - Modeling, representation
- 3.1.8. - Big data (production, storage, transfer)
- 3.2.2. - Knowledge extraction, cleaning
- 3.3.2. - Data mining
- 3.3.3. - Big data analysis
- 3.4.1. - Supervised learning
- 3.4.2. - Unsupervised learning
- 3.4.3. - Reinforcement learning
- 3.4.4. - Optimization and learning
- 3.4.5. - Bayesian methods
- 3.4.6. - Neural networks
- 3.4.7. - Kernel methods
- 3.4.8. - Deep learning
- 5.3.3. - Pattern recognition
- 6.2.4. - Statistical methods
- 6.2.6. - Optimization

Other Research Topics and Application Domains:

- 1.1.10. - Mathematical biology
- 1.1.5. - Genetics
- 1.1.6. - Genomics
- 1.1.9. - Bioinformatics
- 9.4.2. - Mathematics

1. Members

Research Scientists

Kevin Bleakley [Inria, Researcher]
Gilles Celeux [Inria, Senior Researcher]

Faculty Members

Pascal Massart [Team leader, Univ. Paris XI, Professor]
Julie Josse [Agro Rennes, Associate Professor, from Sep 2015]
Christine Keribin [Univ. Paris XI, Associate Professor]
Patrick Pamphile [Univ. Paris XI, Associate Professor]
Jean-Michel Poggi [Univ. Paris V, Professor, HdR]
Yves Rozenholc [Univ. Paris V, Associate Professor, until Aug 2015, HdR]
Claire Lacour [Univ. Paris XI, Associate Professor]
Erwan Le Pennec [Ecole Polytechnique,Professor]

Engineers

Benjamin Auder [CNRS, Researcher]
Yves Misiti [CNRS, until Feb 2015]
Yi Liu [until Sep 2015]
Jonas Renault [Inria, from Oct 2015]

PhD Students

Neska El Haouij [Inria, from Oct 2015]
Emilie Devijver [Univ. Paris XI, until Sep 2015]
Melina Gallopin [Univ. Paris XI]
Jana Kalawoun [Univ. Paris XI]
Claire Brecheteau [Univ. Paris XI, from Oct 2015]
Jeanne Nguyen [Univ. Paris XI, from Oct 2015]
Valerie Robert [Univ. Paris XI]
Solenne Thivin [Thales, until Oct 2015]
Vincent Thouvenot [EDF]
Yann Vasseur [Univ. Paris XI]

Administrative Assistant

Olga Mwana Mobulakani [Inria]

Others

Ignacio Solis Meza [Inria, until May 2015]
Yves Auffray [Dassault Aviation]
Serge Cohen [Ipanema]
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 aims 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, curve classification, phylogenetic analysis and classification in genetics. New developments in SELECT activities are concerned with applications in biostatistics (statistical analysis of medical images) and population genetics.

3. Research Program

3.1. General presentation

From applications we treat on a day-to-day basis, we have learned that some assumptions 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 makes asymptotic analyses breakdown. An important aim of SELECT is to propose model selection criteria which take such practical constraints into account.

3.2. A nonasymptotic view of model selection

An important goal 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 this, and lead to data-driven penalty choice strategies. A major research direction for SELECT consists of deepening the analysis of data-driven penalties, both from the theoretical and practical points of view. There is no universal way of calibrating penalties, but there are several different general ideas that we aim to develop, including heuristics derived from Gaussian theory, special strategies for variable selection, and 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 can avoid or overcome certain theoretical difficulties, and 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 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 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 methodology to address them. Many of our applications involve contracts with industrial partners, e.g., in reliability, although we also have several academic collaborations, e.g., in genetics and image analysis.

4.2. Curve classification

The field of classification for complex data such as curves, functions, spectra and time series, is an important problem in current research. Standard data analysis questions are being looked into anew, in order to define novel strategies that take the functional nature of such data into account. Functional data analysis addresses a variety of applied problems, including longitudinal studies, analysis of fMRI data, and spectral calibration.

We are focused in particular on unsupervised classification. In addition to standard questions such as the choice of the number of clusters, the norm for measuring the distance between two observations, and vectors for representing clusters, we must also address a major computational problem: the functional nature of the data, which requires new approaches.

4.3. Computer experiments and reliability

For several years now, SELECT has collaborated with the EDF-DER *Maintenance des Risques Industriels* group. One important theme involves 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, a high-technology group (Aerospace propulsion, Aicraft equipment, Defense Security, Communications).

Moreover, a collaboration has begun with Dassault Aviation on the modal analysis of mechanical structures, which aims to identify the vibration behavior of structures under dynamic excitation. From the algorithmic point of view, modal analysis amounts to estimation in parametric models on the basis of measured excitations and structural response data. In literature and existing implementations, the model selection problem associated with this estimation is currently treated by a rather weighty and heuristic procedure. In the context of our own research, model selection via penalization methods are to be tested on this model selection problem.

4.4. Dynamic contrast enhanced imaging

Yves Rozenholc was with SELECT until Septemnber 2015, and introduced research for quantifying tumor microcirculation to monitor treatments in cancer. Dynamic Contrast Enhanced (DCE) imaging provides information on the properties of vascular networks. 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 has developed several tools for denoising and clustering the dynamics found in DCE imaging sequences, and testing equality of survival functions coming from two DCE imaging sequences.

4.5. Analysis of genomic data

For many years now, 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 has started a thesis co-supervised 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.

SELECT collaborates with Anavaj Sakuntabhai and Benno Schwikowski (Pasteur Institute) on prediction of dengue fever severity from high-dimensional gene expression data. One project involves using/finding new and computationally efficient methods (e.g., 2d isotonic regression, lasso regression) for predicting dengue severity. Due to the high-dimensional nature of the data and low-dimensional nature of the number of individuals, false discovery rate (FDR) methods are used to provide statistical justification of results. A second project involves statistical meta-analysis of newly collected dengue gene expression data along with recently published data sets from other groups.

SELECT is involved in the ANR “jeunes chercheurs” MixStatSeq directed by Cathy Maugis (INSA Toulouse), which is concerned with statistical analysis and clustering of RNASeq genomics data.

4.6. Pharmacovigilance

A collaboration has been 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 goal is to detect, as soon as possible, potential associations between certain drugs and adverse effects, which appeared after the authorized marketing of these drugs. Instead of working on aggregate data (contingency table) like is usually the case, the approach developed aims to deal with individual’s data, which perhaps gives more information. Valerie Robert has begun a thesis co-supervised by Gilles Celeux and Christine Keribin on this topic, which should enable the development of a new model-based clustering method, inspired by latent block models.

4.7. Spectroscopic imaging analysis of ancient materials

Ancient materials, encountered in archaeology and paleontology are often complex, heterogeneous and poorly characterized before physico-chemical analysis. A popular technique to gather as much physico-chemical information as possible, is spectro-microscopy or spectral imaging, where a full spectra, made of more than a 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 requires the combination of an "image" approach with a "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 datasets. This technique enables 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], Benjamin Auder, Jonas Renault.

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 MODeling) 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 components, some of them favoring either a cluster analysis or a discriminant analysis point of view, are included. Many Gaussian models for continuous variables and multinomial models for discrete variable are included. Written in C++, MIXMOD is interfaced with MATLAB. The software, statistical documentation, and user guide are available here: 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.

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

This year, MIXMOD has received the support of an ADT (MASSICCC) for three years. This ADT MASSICCC has been obtained conjointly with the MODAL team (Inria Lille). This year, an engineer, Jonas Renault, has been appointed for two years. He is in charge of developing a web version of MIXMOD.

5.2. BLOCKCLUSTER software

Participants: Gilles Celeux, Christine Keribin.

Mixture model, Block cluster analysis, Blockcluster is software devoted to model-based block clustering. It is developed in partnership with the MODAL team (Inria Lille). This year, some major bugs have been fixed, and the Bayesian point of view has been reinforced by including Gibbs sampling for binary and categorial data. This Gibbs sampler, coupled with the variational Bayes algorithm, provides solutions which are more stable and less dependent on the initial values of the algorithm. An exact expression of the ICL criterion has been provided. This non-asymptotic criterion appears to be more relevant than the BIC-like approximation of ICL.

Vincent Brault, Christine Keribin and Mahindra Mariadassou have shown the consistency and asymptotic normality of the maximum likelihood and variational estimators in stochastic or latent block models.

6. New Results

6.1. Model selection in Regression and Classification

Participants: Gilles Celeux, Serge Cohen, Erwan Le Pennec, Pascal Massart, Kevin Bleakley.

The well-documented and consistent variable selection procedure in model-based cluster analysis and classification that Cathy Maugis (INSA Toulouse) designed during her PhD thesis in SELECT, makes use of stepwise algorithms which are painfully slow in high dimension. 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 ranking to select variables, they avoid the combinatory problem of stepwise procedures. After tests on challenging simulated and real data sets, their algorithm has shown encouraging performance. Moreover, the possibility to sort the variables with their marginal likelihoods is under study. The first results are encouraging, and this approach requires no regularization hyperparameters, and is much more rapid.

In collaboration with Jean-Michel Marin (Université de Montpellier) and Olivier Gascuel (LIRMM), Gilles Celeux has continued research aiming to select a short list of models rather a single model. This short list 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. Furthermore, the Kullback-Leibler distances at hand are estimated through nonparametric and parametric bootstrap procedures. Different strategies are compared through numerical experiments on simulated and real data sets.

Emilie Devijver, Yannig Goude and Jean-Michel Poggi have proposed a new methodology for customer segmentation, in the context of load profiles in energy consumption. The method is based on high-dimensional regression models which perform clustering and model selection at the same time. They have focused on uncovering classes corresponding to different regression models ,and compute clustering and model identification in each cluster simultaneously. They have shown the feasibility of the approach on a real data set of Irish customers.

Emilie Devijver has studied a dimension-reduction method for finite mixtures of multivariate response regression models in high-dimension. The size of the response and the number of predictors may exceed the sample size. She considers jointly predictor selection and rank reduction to obtain lower-dimensional approximations of parameter matrices. A penalty, for which the model selected by penalized likelihood satisfies an oracle inequality, is given.

The detection of change-points in a spatially or time-ordered data sequence is an important problem in many fields such as genetics and finance. Kevin Bleakley, with Gérard Biau (LSTA, Paris 6 University) and David Mason (University of Delaware), has found asymptotic distributions of statistics used to detect change-points, and developed methods to provide stopping criteria (model selection) for the number of change-points found.

6.2. Statistical learning methodology and theory

Participants: Gilles Celeux, Christine Keribin, Erwan Le Pennec, Michel Prenat, Solenne Thivin, Kevin Bleakley.

Gilles Celeux has started a collaboration with Jean-Patrick Baudry on strategies to avoid traps in the EM algorithm in mixture analysis. They analyze the effect of spurious local maximizers, and regularized algorithms to avoid such solutions. They show the link that exists between the degree of regularization and slope heuristics. Moreover, their strategy to initiate the EM algorithm, embedding the solution with K components and the starting position with $K + 1$ components to avoid suboptimal solutions, has been proved to be efficient, and is extended to a more complex framework of latent block models.

In the context of algorithms that depend on distributed computing and collaborative inference, Kevin Bleakley, with with Gérard Biau (LSTA, Paris 6) and Benoït Cadre (ENS Rennes), have proposed a collaborative framework that aims to estimate the unknown mean θ of a random variable X. In the model they present, a certain number of calculation units, distributed across a communication network represented by a graph, participate in the estimation of θ by sequentially receiving independent data from X while exchanging messages via a stochastic matrix A defined over the graph. They give precise conditions on the matrix A under which the statistical precision of the individual units is comparable to that of a (gold standard) virtual centralized estimate, even though each unit does not have access to all of the data.

6.3. Reliability

Participants: Yves Auffray, Gilles Celeux, Florence Ducros, Patrick Pamphile, Jana Kalawoun.

Since June 2015, in the framework of a CIFRE convention with Nexter, Florence Ducros has commenced a thesis on the modeling of aging of vehicles, supervised by Gilles Celeux and Patrick Pamphile. This thesis should lead to designing an efficient maintenance strategy according to vehicle use profiles. It will involve the estimation of mixtures and competing risk models in a highly censored setting.

Janan Kalawoun has defended her thesis supervised by Gilles Celeux, Patrick Pamphile and Maxime Montaru (CEA) on 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 battery dynamics in embedded applications (e.g., electric vehicles). The SoC is modeled by a switching Markov state-space model. Parameters are estimated by combining the EM algorithm and particle filter methods. The model is validated using real-world electric vehicle data. This model has been proved to be highly superior to a simple state space model. The optimal number of battery modes is then identified, using model selection criteria such as AIC and BIC, which has proved to be superior to cross-validation in this particular context.

In the framework of a study on the dispatch availability of Dassault Aviation business jets, Yves Auffray and Gilles Celeux have contributed to methodology aiming to discover the root causes of reliability flaws.

6.4. Statistical analysis of genomic data

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

Mélina Gallopin defended her thesis supervised by Gilles Celeux, Florence Jaffrezic and Andrea Rau (INRA, animal genetics department), This thesis was concerned with modeling and model selection in the analysis of RNA-seq data. Its highlights are the following:

- Presentation of a model selection criterion for model-based clustering of annotated gene expression data. This criterion is an ICL-like criterion taking into account annotation.
- An objective comparison of discrete and continuous modeling after transformations for RNA-seq data based on a comparison of likelihoods (possibly penalized) of the possible models.
- A block diagonal covariance selection method for high dimensional Gaussian graphical models. This non-asymptotic model selection procedure is supported by strong theoretical guarantees, based on an oracle inequality and a minimax lower bound. This work was in collaboration with Emilie Devijver.

The subject of Yann Vasseur's 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*. To that purpose, a transciptome dataset with a similar number of TFs and statistical units is available. The first aim consists of reducing the dimension of the network to avoid high-dimensional difficulties. Representing this network with a Gaussian graphical model, the following procedure has been defined:

1. *Selection step*: choose the set of TF regulators (supports) of each TF.
2. *Classification step*: deduce co-factors groups (TFs with similar 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 to the dataset. The study of stability and parameter calibration of these methods is in progress. The TFs are clustered with the Latent Block Model in a number of co-factor groups, selected with BIC or the exact ICL criterion.

In a collaboration with Marie-Laure Martin-Magniette, Cathy Maugis and Andrea Rau, Gilles Celeux has studied gene expression obtained from high-throughput sequencing technology. The focus is on the question of clustering gene expression profiles as a means to discover groups of co-expressed genes. A Poisson mixture model is proposed, using a rigorous framework for parameter estimation as well as for 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 and serial analysis of gene expression data. The proposed method is implemented in the open-source R package `HTSCluster`, available on CRAN. It can now be compared with Gaussian mixtures obtained after relevant data transformations.

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 have started research concerning the detection of associations between drugs and adverse events in the framework of the PhD of Valerie Robert. At first, this team developed a model-based clustering inspired by latent block models, which consists of co-clustering rows and columns of two binary tables, imposing the same row ranking. This enables it to highlight subgroups of individuals sharing the same drug profile, and subgroups of adverse effects and drugs with strong interactions. Furthermore, some sufficient conditions are provided to obtain the identifiability of the model, and some results are shown for simulated data. This year, the exact ICL criterion has been extended to this double block latent model. Moreover, the possible added value of this model, compared with standard contingency table analysis, is currently under scrutiny.

7. Bilateral Contracts and Grants with Industry

7.1. Contract with SNECMA

Participants: Gilles Celeux, Florence Ducros, Patrick Pamphile.

SELECT has a contract with Nexter regarding modeling the reliability of vehicles.

SELECT works with the CEA on statistical modeling for battery state of charge.

Contract with AirNormand: Mixtures of experts for PM10 forecasting, and stability of kriging procedures.

Contract with EDF: Curve clustering and disaggregation of the load forecasting

8. Partnerships and Cooperations

8.1. Regional Initiatives

Pascal Massart co-organizes a working group at ENS (Ulm) on statistical learning.

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

8.2. National Initiatives

8.2.1. ANR

SELECT is part of the ANR funded 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 Seattle (USA).

8.4. International Research Visitors

8.4.1. Visits to International Teams

8.4.1.1. Research stays abroad

Jean-Michel Poggi visited Anestis Antoniadis at the University of Cape Town (South Africa), Department of Statistical Sciences, 16-26 February 2015

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

Jean-Michel Poggi:

- Organization of the session: Wavelet Methods in Statistics, at the 8th International Conference of the ERCIM WG on Computational and Methodological Statistics, London, 12-14 December 2015.
- Organization of two sessions on MOOCs, ENBIS 2015, 6-10 Sept 2015, Prague. 1) Presentation session on MOOCs, 2) Realization of MOOCs – technology, content and funding opportunities.
- Organisation of the conference: “MOOC et formation continue en statistique”, 3 mars 2015, IHP, Paris.

9.1.1.2. Member of the organizing committees

Gilles Celeux is one of the co-organizers of the international working group on model-based clustering. This year this workshop took place in Seattle (USA).

9.1.2. Journal

9.1.2.1. Member of the editorial boards

Gilles Celeux is Editor-in-Chief of the *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*. He is also editor (with A. Antoniadis, X. Brossat) of a Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension, Springer 2015.

9.1.2.2. Reviewer - Reviewing activities

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

9.1.3. Invited talks

The members of the team have given many invited talks on their research in the course of 2015.

9.1.4. Leadership within the scientific community

Jean-Michel Poggi is:

- Vice-President ENBIS (European Network for Business and Industrial Statistics), 2015-18
- Vice-President FENStatS (Federation of European National Statistical Societies) since 2012
- Council Member of the ISI (2015-19)
- Member of the Board of Directors of the ERS of IASC (since 2014)

9.1.5. Scientific expertise

Jean-Michel Poggi is member of the EMS Committee for Applied Mathematics (since 2014).

9.1.6. Research administration

Jean-Michel Poggi is the president of ECAS (European Courses in Advanced Statistics) since 2015

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

SELECT members teach various courses at several different universities, and in particular the Master 2 "Modélisation stochastique et statistique" of University Paris-Sud.

9.2.2. Supervision

PhD: Jana Kalawoun, Modélisation statistique de l'état de charge des batteries électriques, Université Paris-Sud, November 2015, Gilles Celeux and Patrick Pamphile

PhD: Mélina Gallopin, Classification et inférence de réseaux pour les données RNA-seq, Université Paris-Sud, December 2015, Gilles Celeux with Andrea Rau and Florence Jaffrezic (INRA)

PhD: Émilie Devivjer, Modèles de mélange pour la régression en grande dimension, application aux données fonctionnelles, Université Paris-Sud, July 2015, Pascal Massart and Jean-Michel Poggi

PhD: Solenne Thivin, Détection automatique d'anomalies sur fonds complexes pour des images ou séquences d'images, Université Paris-Sud, December 2015, Erwan Le Pennec

PhD: Vincent Thouvenot, Estimation et sélection pour les modèles additifs et application à la prévision de la consommation électrique, December 2015, Jean-Michel Poggi and Anestis Antoniadis (Univ. Joseph Fourier, Grenoble)

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

PhD in progress: Yann Vasseur, 2013, Gilles Celeux and 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: Florence Ducros, 2015, Gilles Celeux and Patrick Pamphile

PhD in progress: Claire Brecheteau, 2015, Pascal Massart

PhD in progress: Jeanne Nguyen, 2015, Claire Lacour

9.3. Popularization

Emilie Devijver:

- Organisation of a spring school for high school students about probability, Pristina, Kosovo
- Organisation of the "Séminaire de Vulgarisation des Doctorants" at Université Paris Sud
- Several talks in high schools to give tools to students to understand conferences: "Un texte, un mathématicien" organized at BNF.

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] 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", 2015, 14 p. , https://hal.inria.fr/hal-01108795

[2] J.-P. BAUDRY, G. CELEUX. *EM for mixtures-Initialization requires special care*, in "Statistics and Computing", 2015, https://hal.inria.fr/hal-01256833

[3] E. DEVIJVER. *An ℓ 1 -oracle inequality for the Lasso in finite mixture of multivariate Gaussian regression models*, in "ESAIM: Probability and Statistics", December 2015, https://hal.inria.fr/hal-01075338

[4] E. DEVIJVER. *Finite mixture regression: a sparse variable selection by model selection for clustering*, in "Electronic Journal of Statistics", December 2015, 20 pages, https://hal.archives-ouvertes.fr/hal-01060079

[5] M. GALLOPIN, G. CELEUX, F. JAFFREZIC, A. RAU. *A model selection criterion for model-based clustering of annotated gene expression data*, in "Statistical Applications in Genetics and Molecular Biology", 2015, https://hal.inria.fr/hal-01256765

[6] M. GALLOPIN, G. CELEUX, F. JAFFRÉZIC, A. RAU. *A model selection criterion for model-based clustering of annotated gene expression data*, in "Statistical Applications in Genetics and Molecular Biology", January 2015, vol. 14, n^{o} 5 [*DOI :* 10.1515/SAGMB-2014-0095], https://hal.inria.fr/hal-01255908

[7] C. HERNANDEZ, C. KERIBIN, P. DROBINSKI, S. TURQUETY. *Statistical modelling of wildfire size and intensity: a step toward meteorological forecasting of summer extreme fire risk*, in "Annales Geophysicae", 2015, vol. 33, n^{o} 12, pp. 1495-1506 [*DOI :* 10.5194/ANGEO-33-1495-2015], http://hal.upmc.fr/hal-01260501

[8] C. KERIBIN, V. BRAULT, G. CELEUX, G. GOVAERT. *Estimation and selection for the latent block model on categorical data*, in "Statistics and Computing", 2015, vol. 25, 16 p. , https://hal.inria.fr/hal-01256840

[9] 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", 2015, forthcoming, https://hal.archives-ouvertes.fr/hal-00919486

[10] A. RAU, C. MAUGIS-RABUSSEAU, M.-L. MAGNIETTE, G. CELEUX. *Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models*, in "Bioinformatics", 2015, vol. 31, n^{o} 9, pp. 1420-1427 [*DOI :* 10.1093/BIOINFORMATICS/BTU845], https://hal.archives-ouvertes.fr/hal-01108821

Articles in National Peer-Reviewed Journals

[11] G. CELEUX, V. ROBERT. *Towards an objective team efficiency rate in basketball*, in "Journal de la Socie´te´ Française de Statistique", 2015, vol. 156, n^{o} 2, 19 p. , https://hal.inria.fr/hal-01020295

Invited Conferences

[12] C. KERIBIN. *Choix de modèles quand la vraisemblance est incalculable*, in "47èmes Journées de Statistique de la SFdS", Lille, France, June 2015, https://hal.inria.fr/hal-01260761

International Conferences with Proceedings

[13] J. KALAWOUN, P. PAMPHILE, G. CELEUX, K. BILETSKA, M. MONTARU. *Estimation of the battery state of charge: a switching Markov state-space model*, in "EUSIPCO'2015", Nice, France, August 2015, 5 p. , https://hal.archives-ouvertes.fr/hal-01168344

[14] J. KALAWOUN, P. PAMPHILE, G. CELEUX. *Identifiability of a Switching Markov State-Space Model*, in "Gretsi 2015", Lyon, France, September 2015, 4 p. , https://hal.archives-ouvertes.fr/hal-01168323

[15] Y. LIU, C. KERIBIN, T. POPOVA, Y. ROZENHOLC. *Statistical Estimation of Genomic Tumoral Alterations*, in "47èmes Journées de Statistique de la SFdS", Lille, France, June 2015, https://hal.inria.fr/hal-01260716

Conferences without Proceedings

[16] M. GALLOPIN, E. DEVIJVER. *Optimal Block Diagonal Covariance Matrices in Large Scale Gaussian Graphical Models*, in "StatMathAppli 2015", Fréjus, France, August 2015, https://hal.inria.fr/hal-01256226

[17] M. GALLOPIN, A. RAU, G. CELEUX, F. JAFFRÉZIC. *Transformation des données et comparaison de modèles pour la classification des données RNA-seq*, in "47èmes Journées de Statistique de la SFdS", Lille, France, June 2015, https://hal.inria.fr/hal-01200672

[18] V. ROBERT, G. CELEUX, C. KERIBIN. *Un modèle statistique pour la pharmacovigilance*, in "47èmes Journées de Statistique de la SFdS", Lille, France, June 2015, https://hal.inria.fr/hal-01255701

Research Reports

[19] F. CHAZAL, P. MASSART, B. MICHEL. *Rates of convergence for robust geometric inference*, Inria, March 2015, https://hal.inria.fr/hal-01232197

Other Publications

[20] J.-P. BAUDRY, G. CELEUX. *EM for mixtures - Initialization requires special care*, February 2015, working paper or preprint, https://hal.inria.fr/hal-01113242

[21] G. BIAU, K. BLEAKLEY, B. CADRE. *The Statistical Performance of Collaborative Inference*, July 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01170254

[22] G. BIAU, K. BLEAKLEY, D. MASON. *Long signal change-point detection*, April 2015, working paper or preprint, https://hal.inria.fr/hal-01140119

[23] L. BIRGÉ, N. MAGALHÃES, P. MASSART. *A new V-fold type procedure based on robust tests*, June 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01163771

[24] F. CHAZAL, P. MASSART, B. MICHEL. *Rates of convergence for robust geometric inference*, May 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01157551

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

[26] E. DEVIJVER, M. GALLOPIN. *Block-diagonal covariance selection for high-dimensional Gaussian graphical models*, November 2015, working paper or preprint, https://hal.inria.fr/hal-01227608

[27] E. DEVIJVER, Y. GOUDE, J.-M. POGGI. *Clustering electricity consumers using high-dimensional regression mixture models*, June 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01169324

[28] J. KALAWOUN, P. PAMPHILE. *Signal Processing by Switching Markov State-Space Models: Estimation of the State of Charge of an Electric Battery*, May 2015, working paper or preprint, https://hal.inria.fr/hal-01149641

[29] C. LACOUR, P. MASSART. *Minimal penalty for Goldenshluger-Lepski method*, January 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01121989

[30] A. RAU, M. GALLOPIN, F. JAFFREZIC, G. CELEUX. *ICAL*, 2015, Software, https://hal.archives-ouvertes.fr/hal-01194145