IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2018

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

1. Team, Visitors, External Collaborators **1**
2. Overall Objectives **2**
3. Research Program **2**
3.1. General presentation 2
3.2. A nonasymptotic view of model selection 2
3.3. Taking into account the modeling purpose in model selection 2
3.4. Bayesian model selection 3
4. Application Domains **3**
4.1. Introduction 3
4.2. Curve classification 3
4.3. Computer experiments and reliability 3
4.4. Analysis of genomic data 3
4.5. Pharmacovigilance 4
4.6. Spectroscopic imaging analysis of ancient materials 4
5. New Software and Platforms **4**
5.1. BlockCluster 4
5.2. MASSICCC 4
5.3. Mixmod 5
6. New Results **5**
6.1. Model selection in Regression and Classification 5
6.2. Estimator selection and statistical tests 6
6.3. Statistical learning methodology and theory 6
6.4. Statistical analysis of genomic data 7
6.5. Reliability 7
6.6. Dynamical systems 8
6.7. Soccer forecasting 8
6.8. Electricity load forecasting and clustering 8
7. Bilateral Contracts and Grants with Industry **9**
8. Partnerships and Cooperations **9**
8.1. Regional Initiatives 9
8.2. National Initiatives 9
8.3. International Initiatives 9
8.4. International Research Visitors 9
9. Dissemination **9**
9.1. Promoting Scientific Activities 9
9.1.1. Scientific Events Organisation 9
9.1.1.1. General Chair, Scientific Chair 9
9.1.1.2. Member of the Organizing Committees 10
9.1.2. Scientific Events Selection 10
9.1.3. Journal 10
9.1.3.1. Member of the Editorial Boards 10
9.1.3.2. Reviewer - Reviewing Activities 10
9.1.4. Invited Talks 10
9.1.5. Leadership within the Scientific Community 10
9.1.6. Scientific Expertise 10
9.1.7. Research Administration 10
9.2. Teaching - Supervision - Juries 10
9.2.1. Teaching 10
9.2.2. Supervision 11

9.2.3. Juries 11
10. Bibliography **11**

Project-Team SELECT

Creation of the Project-Team: 2007 January 01, end of the Project-Team: 2018 December 31

Keywords:

Computer Science and Digital Science:

A3.1.1. - Modeling, representation
A3.1.8. - Big data (production, storage, transfer)
A3.2.2. - Knowledge extraction, cleaning
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.3. - Reinforcement learning
A3.4.4. - Optimization and learning
A3.4.5. - Bayesian methods
A3.4.6. - Neural networks
A3.4.7. - Kernel methods
A3.4.8. - Deep learning
A5.3.3. - Pattern recognition
A6.2.4. - Statistical methods
A6.2.6. - Optimization

Other Research Topics and Application Domains:

B1.1.4. - Genetics and genomics
B1.1.7. - Bioinformatics
B1.1.8. - Mathematical biology
B9.5.2. - Mathematics

1. Team, Visitors, External Collaborators

Research Scientists

Kevin Bleakley [Inria, Researcher]
Gilles Celeux [Inria, Emeritus]
Matthieu Lerasle [CNRS, Researcher]

Faculty Members

Pascal Massart [Team leader, Univ Paris-Sud, Professor]
Sylvain Arlot [Univ Paris-Sud, Professor]
Christine Keribin [Univ Paris-Sud, Associate Professor]
Patrick Pamphile [Univ Paris-Sud, Associate Professor]
Jean-Michel Poggi [Univ René Descartes, Professor, HDR]

PhD Students

Florence Ducros [Univ Paris-Sud, until Sep 2018]
Neska El Haouij [Univ Paris-Sud, until Sep 2018]
Benjamin Goehry [Univ Paris-Sud]
Hedi Hadiji [Univ Paris-Sud]

Minh Lien Nguyen [Univ Paris-Sud, until Oct 2018]

Technical staff

Benjamin Auder [CNRS]
Christian Poli [Inria]

Administrative Assistant

Maeva Jeannot [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 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 biology.

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 reliability, through a research collaboration with Nexter involving a Cifre convention. This collaboration concerns a lifetime analysis of a vehicle fleet to assess ageing.

Moreover, a collaboration is ongoing 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 being tested on this model selection problem.

4.4. 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.

SELECT collaborates with Anavaj Sakuntabhai and Philippe Dussart (Pasteur Institute) on predicting dengue severity using only low-dimensional clinical data obtained at hospital arrival. Further collaborations are underway in dengue fever and encephalitis with researchers at the Pasteur Institute, including with Jean-David Pommier.

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.

4.5. Pharmacovigilance

A collaboration is ongoing 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.

4.6. 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 clustering problems, taking spatial constraints into account.

5. New Software and Platforms

5.1. BlockCluster

Block Clustering
KEYWORDS: Statistic analysis - Clustering package
SCIENTIFIC DESCRIPTION: Simultaneous clustering of rows and columns, usually designated by biclustering, co-clustering or block clustering, is an important technique in two way data analysis. It consists of estimating a mixture model which takes into account the block clustering problem on both the individual and variables sets. The blockcluster package provides a bridge between the C++ core library and the R statistical computing environment. This package allows to co-cluster binary, contingency, continuous and categorical data-sets. It also provides utility functions to visualize the results. This package may be useful for various applications in fields of Data mining, Information retrieval, Biology, computer vision and many more.
FUNCTIONAL DESCRIPTION: BlockCluster is an R package for co-clustering of binary, contingency and continuous data based on mixture models.

- Participants: Christophe Biernacki, Gilles Celeux, Parmeet Bhatia, Serge Iovleff, Vincent Brault and Vincent Kubicki
- Partner: Université de Technologie de Compiègne
- Contact: Serge Iovleff
- URL: http://cran.r-project.org/web/packages/blockcluster/index.html

5.2. MASSICCC

Massive Clustering with Cloud Computing
KEYWORDS: Statistic analysis - Big data - Machine learning - Web Application

SCIENTIFIC DESCRIPTION: The web application let users use several software packages developed by Inria directly in a web browser. Mixmod is a classification library for continuous and categorical data. MixtComp allows for missing data and a larger choice of data types. BlockCluster is a library for co-clustering of data. When using the web application, the user can first upload a data set, then configure a job using one of the libraries mentioned and start the execution of the job on a cluster. The results are then displayed directly in the browser allowing for rapid understanding and interactive visualisation.
FUNCTIONAL DESCRIPTION: The MASSICCC web application offers a simple and dynamic interface for analysing heterogeneous data with a web browser. Various software packages for statistical analysis are available (Mixmod, MixtComp, BlockCluster) which allow for supervised and supervised classification of large data sets.

- Contact: Christophe Biernacki
- URL: https://massiccc.lille.inria.fr

5.3. Mixmod

Many-purpose software for data mining and statistical learning
KEYWORDS: Data mining - Classification - Mixed data - Data modeling - Big data
FUNCTIONAL DESCRIPTION: Mixmod is a free toolbox for data mining and statistical learning designed for large and highdimensional data sets. Mixmod provides reliable estimation algorithms and relevant model selection criteria.

It has been successfully applied to marketing, credit scoring, epidemiology, genomics and reliability among other domains. Its particularity is to propose a model-based approach leading to a lot of methods for classification and clustering.

Mixmod allows to assess the stability of the results with simple and thorough scores. It provides an easy-to-use graphical user interface (mixmodGUI) and functions for the R (Rmixmod) and Matlab (mixmodForMatlab) environments.

- Participants: Benjamin Auder, Christophe Biernacki, Florent Langrognet, Gérard Govaert, Gilles Celeux, Remi Lebret and Serge Iovleff
- Partners: CNRS - Université Lille 1 - LIFL - Laboratoire Paul Painlevé - HEUDIASYC - LMB
- Contact: Gilles Celeux
- URL: http://www.mixmod.org

6. New Results

6.1. Model selection in Regression and Classification

Participants: Gilles Celeux, Pascal Massart, Sylvain Arlot, Jean-Michel Poggi, Kevin Bleakley.

In collaboration with Damien Garreau, Sylvain Arlot studied the kernel change-point algorithm (KCP) proposed by Arlot, Celisse and Harchaoui (2012), which aims at locating an unknown number of change-points in the distribution of a sequence of independent data taking values in an arbitrary set. The change-points are selected by model selection with a penalized kernel empirical criterion. We provide a non-asymptotic result showing that, with high probability, the KCP procedure retrieves the correct number of change-points, provided that the constant in the penalty is well-chosen; in addition, KCP estimates the change-points location at the optimal rate. As a consequence, when using a characteristic kernel, KCP detects all kinds of change in the distribution (not only changes in the mean or the variance), and it is able to do so for complex structured data (not necessarily in Rd). Most of the analysis is conducted assuming that the kernel is bounded; part of the results can be extended when we only assume a finite second-order moment.

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 dimensions. In order to circumvent this drawback, Gilles Celeux, in collaboration with Mohammed Sedki (Université Paris XI) and Cathy Maugis, have recently submitted an article where variables are sorted 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. The performances on challenging simulated and real data sets are similar to the standard procedure, with a CPU time divided by a factor of more than a hundred.

In collaboration with Benjamin Charlier and Jean-Michel Marin (Université de Montpellier), Gilles Celeux has started research aiming to propose a rapid Bayesian algorithm to estimate simply the mode of a posterior distribution for hidden structure models. This Bayesian procedure is of interest for two reasons. First, it leads to regularised estimation, which is useful for poorly posed problems. Second, it is an interesting alternative to variational approximation.

6.2. Estimator selection and statistical tests

Participant: Sylvain Arlot.

Sylvain Arlot wrote a book chapter about cross-validation in 2018. This text defines all classical cross-validation procedures, and studies their properties for two different goals: estimating the risk of a given estimator, and selecting the best estimator among a given family. For the risk estimation problem, it computes the bias (which can also be corrected) and the variance of cross-validation methods. For estimator selection, it first provides a first-order analysis (based on expectations). Then, it explains how to take into account second-order terms (from variance computations, and by taking into account the usefulness of over-penalization). This allows, in the end, to provide some guidelines for choosing the best cross-validation method for a given learning problem.

6.3. Statistical learning methodology and theory

Participants: Gilles Celeux, Serge Cohen, Christine Keribin, Michel Prenat, Sylvain Arlot, Benjamin Auder, Jean-Michel Poggi, Neska El Haouij, Kevin Bleakley, Matthieu Lerasle.

Sylvain Arlot wrote a book chapter about supervised statistical learning, from the mathematical point of view in 2018. This text describes the general prediction problem and the two key examples of regression and binary classification. Then, it studies two kinds of learning rules: empirical risk minimizers, which naturally lead to convex risks in classification, and local averaging rules, for which a universal consistency result can be obtained. Finally, it identifies the limits of learning in order to underline its challenges. The text ends with some useful probabilistic tools and some exercises.

Gilles Celeux and Serge Cohen have started research in collaboration with Agnès Grimaud (UVSQ) to perform clustering of hyperspectral images which respects spatial constraints. This is a one-class classification problem where distances between spectral images are given by the χ^2 distance, while spatial homogeneity is associated with a single link distance. This year they have developed a hybrid hierarchical clustering procedure in which sub-clusters respecting spatial consistency are constructed. Then, these sub-clusters are merged without taking spatial constraints into account. This strategy leads to a more realistic segmentation of spectral images.

Gilles Celeux continued his collaboration with Jean-Patrick Baudry on model-based clustering. Last year, they started work on assessing model-based clustering methods on cytometry data sets. The interest of these is that they involve combining clustering and classification tasks in a unified framework. This year, this work was completed, and performed well in comparison with state-of-the-art procedures.

Gillies Celeux has continued research on missing data for model-based clustering in collaboration with Christophe Biernacki (Modal team, Inria Lille) and Julie Josse (École Polytechnique). This year, they implemented several algorithms to estimate their logistic model for mixture analysis involving not missing-at-random mixtures.

In the framework of MASSICCC, Benjamin Auder and Gilles Celeux have started research on the graphical representation of model-based clusters. The aim of this is to better-display proximity between clusters.It leads to a simple procedure to represent the proximity between clusters without any additional assumptions.

After having proved the consistency and asymptotic normality of Latent Block Model estimators with V. Brault and M. Mariadassou, Christine Keribin has worked on the behavior of the ICL and BIC model criteria in this model, and in particular on their probable asymptotic equivalence.

Christine Keribin has started a new collaboration with Christophe Biernacki (Inria Modal Team) to study the ability for co-clustering to be a good regularized method for clustering in HD, which was presented at the CMStatistics 2018 conference.

J-M. Poggi (with R. Genuer), published a survey paper dedicated to "Arbres CART et Forêts aléatoires, Importance et sélection de variables", as a book chapter published in: "Apprentissage Statistique et Données Massives" by Technip.

J.-M. Poggi and N. El Haouij (with R. Ghozi, S. Sevestre Ghalila and M. Jaïdane) provide a random forest-based method for the selection of physiological functional variables in order to classify the stress level during real-world driving experience. The contribution of this study is twofold: on the methodological side, it considers physiological signals as functional variables and offers a procedure of data processing and variable selection. On the applied side, the proposed method provides a "blind" procedure of driver's stress level classification that does not depend on the expert-based studies of physiological signals. This work has been published in Statistical Methods & Applications.

J.-M. Poggi and N. El Haouij (with R. Ghozi, S. Sevestre Ghalila and M. Jaïdane provide a system and database to assess driver's attention, called aAffectiveROAD. A paper presenting it has been published in the proceedings of the 33rd ACM Symposium on Applied Computing SAC'18.

6.4. Statistical analysis of genomic data

Participant: Kevin Bleakley.

In collaboration with Benno Schwikowski, Iryna Nikolayeva and Anavaj Sakuntabhai (Pasteur Institute, Paris), Kevin Bleakley worked on using 2-d isotonic regression to predict dengue fever severity at hospital arrival using high-dimensional microarray gene expression data. Important marker genes for dengue severity have been detected, some of which now have been validated in external lab trials, and an article on this was published in the Journal of Infectious Diseases in 2018.

Kevin Bleakley has also collaborated with Inserm/Paris-Saclay researchers at Kremlin-Bicêtre hospital on cyclic transcriptional clocks and renal corticosteroid signaling, and has developed novel statistical tests for detecting synchronous signals. This work was published in the FASEB journal in 2018.

Kevin Bleakley worked as part of a consortium on a crowdsourced Dream Challenge in 2018 on using molecular signatures to predict susceptibility to viral infection. Essentially, many teams of researchers from around the world used machine learning (statistical learning) algorithms to learn on training data then test on unseen real data. In the final stage, methods from several teams were combined to improve overall prediction performance. The article "A crowdsourced analysis to identify ab initio molecular signatures predictive of susceptibility to viral infection" was published in Nature Communications in 2018.

6.5. Reliability

Participants: Gilles Celeux, Florence Ducros, Patrick Pamphile.

From June 2015 until June 2018 when she defended it, in the framework of a CIFRE convention with Nexter, Florence Ducros researched 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. Moreover, warranty cost calculations are made in the context of heterogeneous usages. This required estimations of mixtures and competing risk models in a highly-censored setting.

This year, Patrick Pamphile and Florence Ducros have published an article which proposes a two-component Weibull mixture model for modelling unobserved heterogeneity in heavily censored lifetime data collection. Performance of classical estimation methods (maximum of likelihood, EM, full Bayes and MCMC) are poor due to the high number of parameters and the heavy censoring. Thus, a Bayesian bootstrap method called Bayesian Restoration Maximization, was used. Sampling from the posterior distribution was obtained thanks to an importance sampling technique. Simulation results showed that, even with heavy censoring, BRM is effective both in term of estimate's precision and computation times.

6.6. Dynamical systems

Participant: Sylvain Arlot.

In collaboration with Stefano Marmi and Duccio Papini, Sylvain Arlot proposed a new model for the time evolution of livestock commodities which exhibits endogenous deterministic stochastic behaviour. The model is based on the Yoccoz-Birkeland integral equation, a model first developed for studying the time-evolution of single species with high average fertility, a relatively short mating season and density dependent reproduction rates. This equation is then coupled with a differential equation describing the price of a livestock commodity driven by the unbalance between its demand and supply. At its birth the cattle population is split into two parts: reproducing females and cattle for butchery. The relative amount of the two is determined by the spot price of the meat. We prove the existence of an attractor and we investigate numerically its properties: the strange attractor existing for the original Yoccoz-Birkeland model is persistent but its chaotic behaviour depends also from the price evolution in an essential way.

6.7. Soccer forecasting

Participants: Gilles Celeux, Jean-Louis Foulley.

In collaboration with Jean-Louis Foulley (Montpellier University), Gilles Celeux has proposed a penalty criterion for assessing correct score forecasting in soccer matches. They have defined the subject of a Masters internship for next year to predict scores of soccer matches via Poisson models using maximum likelihood and Bayesian inference.

6.8. Electricity load forecasting and clustering

Participants: Jean-Michel Poggi, Benjamin Auder, Benjamin Goehry.

B. Auder, J-M. Poggi (with J. Cugliari, Y. Goude) are interested in hierarchical time-series for bottom-up forecasting. The idea is to disaggregate the signal in such a way that the sum of disaggregated forecasts improves the direct prediction. The 3-steps strategy defines numerous super-consumers by curve clustering, builds a hierarchy of partitions and selects the best one minimizing a forecast criterion. Using a nonparametric model to handle forecasting, and wavelets to define various notions of similarity between load curves, this disaggregation strategy applied to French individual consumers leads to a gain of 16% in forecast accuracy. Then the upscaling capacity of this strategy facing massive data is explored and different proposals using R are experimented. The proposed solutions to make the algorithm scalable combines data storage, parallel computing and double clustering step to define the super-consumers. This has been published in the journal Energies.

Benjamin Goehry is completing a thesis co-supervised by P. Massart and J-M. Poggi, aiming at extending this scheme by introducing the use of random forests as time series forecasting models adapted to each cluster.

J.-M. Poggi (with J. Cugliari) published in Wiley StatsRef-Statistics Reference Online, a paper entitled Electricity demand forecasting. the focus is on short-term demand forecasting at some aggregate level (e.g., zone or nationwide demands) from data with at least hourly sampled data. The main salient features of the load curve are first highlighted. Some of the common covariates used in the prediction task are also discussed. Then, some basic or now classical methodological approaches for electricity demand forecasting are detailed.

7. Bilateral Contracts and Grants with Industry

7.1. Contract with NEXTER

Participants: Gilles Celeux, Florence Ducros, Patrick Pamphile.

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

8. Partnerships and Cooperations

8.1. Regional Initiatives

Sylvain Arlot and Pascal Massart co-organize a working group at ENS (Ulm) on statistical learning.

8.2. National Initiatives

8.2.1. ANR

SELECT is part of the ANR-funded MixStatSeq.

Sylvain Arlot and Matthieu Lerasle are part of the ANR grant FAST-BIG (Efficient Statistical Testing for high-dimensional Models: application to Brain Imaging and Genetics), which is lead by Bertrand Thirion (Inria Saclay, Parietal).

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 Ann Arbor (USA).

8.4. International Research Visitors

8.4.1. Visits to International Teams

8.4.1.1. Research Stays Abroad

Kevin Bleakley stayed at the Pasteur Institute, Cambodia, while working on several collaborations in dengue fever and encephalitis, from February–March 2018.

Jean-Michel Poggi: Universidad de la República (Montevideo, Uruguay), Facultad de Ingeniería, Instituto de Matemática y Estadística “Prof. Ing. Rafael Laguardia”, 17-28 February 2018.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

Sylvain Arlot co-organized (with Christophe Giraud and Gilles Stoltz) the conference “Deux complices en statistique” (Two days in honor of Pascal Massart and Lucien Birgé), at IHES (Bures-sur-Yvette) and IMO (Orsay).

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 the workshop took place in Ann Arbor, USA.
- Sylvain Arlot is one of the co-organizers of the Junior Conference on Data Science and Engineering at Paris-Saclay (3nd edition in 2018).

9.1.2. Scientific Events Selection

9.1.2.1. Member of the Conference Program Committees

Jean-Michel Poggi was a member of the Scientific Programme Committee, ENBIS 2018, Nancy, 2-6, September 2018, and member of the Scientific Committee of the summer school on Clustering, Data, Analysis And Visualization Of Complex Data, May 21-25, 2018.

9.1.3. Journal

9.1.3.1. 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*.

Sylvain Arlot is associate editor for the Annales de l'Institut Henri Poincaré B, Probability and Statistics.

9.1.3.2. Reviewer - Reviewing Activities

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

9.1.4. Invited Talks

The members of the team have given many invited talks on their research in the course of 2018.

9.1.5. Leadership within the Scientific Community

Jean-Michel Poggi is:

- Vice-President of ECAS (European Courses in Advanced Statistics)
- Council Member of the ISI (2015-19)
- Member of the Board of Directors of the ERS of IASC (since 2014)
- Council member of FENStatS (Federation of European National Statistical Societies)

9.1.6. Scientific Expertise

Jean-Michel Poggi was member of the Box Medal committee for 2018, and in the jury for the Marie-Jeanne Laurent Duhamel prize (SFdS).

9.1.7. Research Administration

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

Sylvain Arlot coordinates (jointly with Marc Schoenauer, Inria Saclay) the math-STIC program of the Labex Mathématique Hadamard.

Christine Keribin is treasurer of the Société Française de Statistique (SFdS).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

SELECT members teach various courses at several different universities, and in particular the Master 2 "Mathématique de l'aléatoire" of Université Paris-Saclay.

9.2.2. Supervision

PhD: Neska El Haouij, 2014, Jean-Michel Poggi, Meriem Jaïdane, Raja Ghozi (ENIT Tunisie), and Sylvie Sevestre-Ghalila (CEA LinkLab). Defended in July 2018.

PhD: Florence Ducros, 2015, Gilles Celeux and Patrick Pamphile. Defended June 2018.

PhD in progress: Claire Brécheteau, 2015, Pascal Massart

PhD in progress: Hedi Hadiji, 2017, Pascal Massart

PhD in progress: Guillaume Maillard, 2016, Sylvain Arlot and Matthieu Lerasle

PhD in progress: Jeanne Nguyen, 2015, Claire Lacour and Vincent Rivoirard (Univ Paris Dauphine)

PhD in progress: Benjamin Goehry, 2015, Pascal Massart and Jean-Michel Poggi

PhD in progress: Tuan-Binh Nguyen, 2018, Sylvain Arlot and Bertrand Thirion

9.2.3. Juries

HDR: Emilie Lebarbier (Pascal Massart, referee; Sylvain Arlot, president)

Members of SELECT have participated in numerous juries during 2018.

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] G. CELEUX, C. MAUGIS-RABUSSEAU, M. SEDKI. *Variable selection in model-based clustering and discriminant analysis with a regularization approach*, in "Advances in Data Analysis and Classification", 2018 [*DOI :* 10.1007/s11634-018-0322-5], https://hal.inria.fr/hal-01053784

[2] F. DUCROS, P. PAMPHILE. *Bayesian estimation of Weibull mixture in heavily censored data setting*, in "Reliability Engineering and System Safety", December 2018, https://hal.archives-ouvertes.fr/hal-01645618

[3] S. FOURATI, A. TALLA, M. MAHMOUDIAN, J. BURKHART, R. KLÉN, R. HENAO, T. YU, Z. AYDIN, K. Y. YEUNG, M. E. AHSEN, R. ALMUGBEL, S. JAHANDIDEH, X. LIANG, T. NORDLING, M. SHIGA, A. STANESCU, R. VOGEL, G. PANDEY, C. CHIU, M. MCCLAIN, C. WOODS, G. GINSBURG, L. ELO, E. TSALIK, L. MANGRAVITE, S. SIEBERTS, K. BLEAKLEY. *A crowdsourced analysis to identify ab initio molecular signatures predictive of susceptibility to viral infection*, in "Nature Communications", December 2018, vol. 9, n° 1, https://hal.inria.fr/hal-01960254

[4] D. GARREAU, S. ARLOT. *Consistent change-point detection with kernels*, in "Electronic journal of statistics ", December 2018, vol. 12, n° 2, pp. 4440-4486, https://arxiv.org/abs/1612.04740 , https://hal.archives-ouvertes.fr/hal-01416704

[5] F. LE BILLAN, L. AMAZIT, K. BLEAKLEY, Q.-Y. XUE, E. PUSSARD, C. LHADJ, P. KOLKHOF, S. VIENGCHAREUN, J. FAGART, M. LOMBES. *Corticosteroid receptors adopt distinct cyclical transcriptional signatures*, in "FASEB Journal", October 2018, vol. 32, n° 10, pp. 5626-5639, https://hal.inria.fr/hal-01960262

[6] I. NIKOLAYEVA, P. BOST, I. CASADEMONT, V. DUONG, F. KOETH, M. PROT, U. CZERWINSKA, S. LY, K. BLEAKLEY, T. CANTAERT, P. DUSSART, P. BUCHY, E. SIMON-LORIERE, A. SAKUNTABHAI, B. SCHWIKOWSKI. *A Blood RNA Signature Detecting Severe Disease in Young Dengue Patients at Hospital Arrival*, in "Journal of Infectious Diseases", June 2018, vol. 217, n° 11, pp. 1690-1698 [*DOI :* 10.1093/INFDIS/JIY086], https://hal.inria.fr/hal-01960247

[7] L. RAYNAL, J.-M. MARIN, P. PUDLO, M. RIBATET, C. ROBERT, A. ESTOUP, O. STEGLE. *ABC random forests for Bayesian parameter inference*, in "Bioinformatics", October 2018, https://hal.archives-ouvertes.fr/hal-01961126

Invited Conferences

[8] C. KERIBIN, G. CELEUX, V. ROBERT. *The Latent Block Model: a useful model for high dimensional data*, in "Mixture models : Theory and applications", Paris, France, June 2018, https://hal.archives-ouvertes.fr/hal-01957710

International Conferences with Proceedings

[9] F. CHAZAL, V. DIVOL. *The density of expected persistence diagrams and its kernel based estimation*, in "Symposium of Computational Geometry (SoCG 2018)", Budapest, Hungary, June 2018, Extended version of a paper to appear in the proceedings of the Symposium of Computational Geometry 2018, https://hal.archives-ouvertes.fr/hal-01716181

Conferences without Proceedings

[10] C. BIERNACKI, B. AUDER, G. CELEUX, J. DEMONT, F. LANGROGNET, V. KUBICKI, C. POLI, J. RENAULT. *MASSICCC: A SaaS Platform for Clustering and Co-Clustering of Mixed Data*, in "Workshop MixStatSeq: “Mixture models: Theory and applications”", Paris, France, June 2018, https://hal.archives-ouvertes.fr/hal-01949175

[11] C. BIERNACKI, G. CELEUX, J. JOSSE, F. LAPORTE. *Dealing with missing data in model-based clustering through a MNAR model*, in "CMStatistics 2018 - 11th International Conference of the ERCIM WG on Computational and Methodological Statistics", Pise, Italy, December 2018, https://hal.archives-ouvertes.fr/hal-01949120

[12] C. KERIBIN, C. BIERNACKI. *Co-clustering: A versatile way to perform clustering in high dimension*, in "The 11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)", Pise, Italy, December 2018, https://hal.archives-ouvertes.fr/hal-01949116

Scientific Books (or Scientific Book chapters)

[13] S. ARLOT. *Cross-validation*, in "Apprentissage statistique et données massives", M. MAUMY-BERTRAND, G. SAPORTA, C. THOMAS-AGNAN (editors), Editions Technip, May 2018, https://arxiv.org/abs/1703.03167 , https://hal.archives-ouvertes.fr/hal-01485508

[14] S. ARLOT. *Tutorial on statistical learning*, in "Apprentissage statistique et données massives", M. MAUMY-BERTRAND, G. SAPORTA, C. THOMAS-AGNAN (editors), Editions Technip, May 2018, https://hal.archives-ouvertes.fr/hal-01485506

[15] G. CELEUX, S. FRÜHWIRTH-SCHNATTER, C. ROBERT. *Model Selection for Mixture Models-Perspectives and Strategies*, in "Handbook of Mixture Analysis", CRC Press, December 2018, https://hal.archives-ouvertes.fr/hal-01961077

[16] S. FRÜHWIRTH-SCHNATTER, G. CELEUX, C. P. ROBERT. *Handbook of Mixture Analysis*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, November 2018, pp. 1-536, https://hal.inria.fr/hal-01928103

[17] C. ROBERT, G. CELEUX, K. KAMARY, G. MALSINER-WALLI, J.-M. MARIN. *Computational Solutions for Bayesian Inference in Mixture Models*, in "Handbook of Mixture Analysis", CRC Press, December 2018, https://arxiv.org/abs/1812.07240 , https://hal.archives-ouvertes.fr/hal-01961038

Research Reports

[18] Y. AUFFRAY. *A variational interpretation of classification EM*, Dassault Aviation, September 2018, https://hal.inria.fr/hal-01882980

[19] M. LERASLE, Z. SZABÓ, T. MATHIEU, G. LECUÉ. *MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means*, CNRS / LMO - Laboratoire de Mathématiques d'Orsay, Orsay ; Ecole Polytechnique (Palaiseau, France) ; Laboratoire de Mathématiques d'Orsay ; ENSAE ParisTech ; Inria Saclay, équipe SELECT, October 2018, https://hal.archives-ouvertes.fr/hal-01705881

Other Publications

[20] E. AAMARI, J. KIM, F. CHAZAL, B. MICHEL, A. RINALDO, L. WASSERMAN. *Estimating the Reach of a Manifold*, January 2018, https://arxiv.org/abs/1705.04565 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01521955

[21] S. ARLOT, S. MARMI, D. PAPINI. *Coupling the Yoccoz-Birkeland population model with price dynamics: chaotic livestock commodities market cycles*, May 2018, https://arxiv.org/abs/1803.05404 - 26 pages, 19 figures, https://hal.archives-ouvertes.fr/hal-01812794

[22] M. BAELDE, C. BIERNACKI, R. GREFF. *Real-Time Monophonic and Polyphonic Audio Classification from Power Spectra*, July 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01834221

[23] C. BRÉCHETEAU, A. FISCHER, C. LEVRARD. *Robust Bregman Clustering*, December 2018, https://arxiv.org/abs/1812.04356 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01948051

[24] G. CHINOT, G. LECUÉ, M. LERASLE. *Statistical learning with Lipschitz and convex loss functions*, November 2018, https://arxiv.org/abs/1810.01090 - working paper or preprint [*DOI :* 10.01090], https://hal.archives-ouvertes.fr/hal-01923033

[25] A. HAVET, M. LERASLE, É. MOULINES. *Density estimation for RWRE*, June 2018, https://arxiv.org/abs/1806.05839 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01815990

[26] S. LE CORFF, M. LERASLE, E. VERNET. *A Bayesian nonparametric approach for generalized Bradley-Terry models in random environment*, August 2018, https://arxiv.org/abs/1808.08104 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01860352

[27] G. LECUÉ, M. LERASLE. *Robust machine learning by median-of-means : theory and practice*, November 2018, https://arxiv.org/abs/1711.10306 - 48 pages, 6 figures, https://hal.archives-ouvertes.fr/hal-01923036

[28] G. LECUÉ, M. LERASLE, T. MATHIEU. *Robust classification via MOM minimization*, November 2018, https://arxiv.org/abs/1808.03106 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01923035

[29] Y. LIU, C. KERIBIN, T. POPOVA, Y. ROZENHOLC. *Statistical estimation of genomic alterations of tumors*, January 2018, working paper or preprint, https://hal.inria.fr/hal-01688885

[30] M.-L. J. NGUYEN. *Nonparametric method for sparse conditional density estimation in moderately large dimensions*, January 2018, https://arxiv.org/abs/1801.06477 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01688664

[31] C. ROBERT, G. CELEUX, J. JEWSON, J. JOSSE, J.-M. MARIN, C. P. ROBERT. *Some discussions on the Read Paper "Beyond subjective and objective in statistics" by A. Gelman and C. Hennig*, January 2019, https://arxiv.org/abs/1705.03727 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01968779