IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2017

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:

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:

B1.1.5. - Genetics
B1.1.6. - Genomics
B1.1.9. - Bioinformatics
B1.1.10. - Mathematical biology
B9.4.2. - Mathematics

1. Personnel

Research Scientists

Sylvain Arlot [CNRS, Researcher]
Benjamin Auder [CNRS, Researcher]
Kevin Bleakley [Inria, Researcher]
Gilles Celeux [Inria, Emeritus, HDR]
Matthieu Lerasle [CNRS, Researcher]

Faculty Members

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

Post-Doctoral Fellow

Kaniav Kamary [Inria, until Jul 2017]

PhD Students

Benjamin Goehry [Univ Paris-Sud]
Valérie Robert [Univ Paris-Sud, until Aug 2017]
Yann Vasseur [Univ Paris-Sud, until Aug 2017]
Neska El Haouij [Univ Paris-Sud]
Hedi Hadiji [Univ Paris-Sud]
Minh Lien Nguyen [Univ Paris-Sud]
Florence Ducros [Univ Paris-Sud]

Technical staff

Josselin Demont [Inria, until Jan 2017]
Jonas Renault [Inria, until Sep 2017]
Christian Poli [Inria]

Administrative Assistant

Olga Mwana Mobulakani [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.

Yann Vasseur has completed 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. For this work, Yann Vasseur dealt with high-dimensional ill-posed problems where the number of variable was almost equal to the number of observations. He designed heuristic tools using regularized regression methods to circumvent this difficulty.

SELECT collaborates with Anavaj Sakuntabhai and Philippe Dussart (Pasteur Institute) on predicting dengue severity using only low-dimensional clinical data obtained at hospital arrival. An ongoing project also involves statistical meta-analysis of newly collected dengue gene expression data along with recently published data sets from other groups. Further collaborations are underway in dengue fever and encephalitis with researchers at the Pasteur Institute.

SELECT collaborates with Inserm/Paris-Saclay researchers at Kremlin-Bicêtre hospital on cyclic transcriptional clocks and renal corticosteroid signaling, developing statistical tests for synchronous signals.

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. Valerie Robert has completed a thesis co-supervised by Gilles Celeux and Christine Keribin on this topic, which involved the development of a new model-based clustering method, inspired by latent block models. Morever, she has defined new tools to estimate and assess the block clustering involved in these models.

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 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. 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: Jonas Renault
- URL: https://massiccc.lille.inria.fr

5.3. Mixmod

Many-purpose software for data mining and statistical learning
KEYWORDS: Data modeling - Mixed data - Classification - Data mining - 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.

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 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.

6.2. Estimator selection and statistical tests

Participants: Sylvain Arlot, Matthieu Lerasle.

G. Maillard, S. Arlot and M. Lerasle studied a method mixing cross-validation with aggregation, called aggregated hold-out (Agghoo), which is already used by several practitioners. Agghoo can also be related to bagging. According to numerical experiments, Agghoo can improve significantly cross-validation's prediction error, at the same computational cost; this makes it very promising as a general-purpose tool for prediction. This work provides the first theoretical guarantees on Agghoo, in the supervised classification setting, ensuring that one can use it safely: at worst, Agghoo performs like hold-out, up to a constant factor. A non-asymptotic oracle inequality is also proved, in binary classification under the margin condition, which is sharp enough to get (fast) minimax rates.

With G. Lecué, Matthieu Lerasle working on “learning from MOM's principles”, showing that a recent procedure by Lugosi and Mendelson can be derived by applying Le Cam's “estimation from tests” procedure to MOM's tests. They also showed some robustness properties of these estimators, proving that the rates of convergence of this estimator are not downgraded even if some “outliers” have corrupted the dataset, and the other data have only first and second moments equal to that of the targeted probability distribution.

6.3. Statistical learning methodology and theory

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

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.

Gilles Celeux continued his collaboration with Jean-Patrick Baudry on model-based clustering. This 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.

Gillies Celeux and Julie Josse have started research on missing data for model-based clustering in collaboration with Christophe Biernacki (Modal, Inria Lille). This year, they have proposed a 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.

For a long time unsolved, the consistency and asymptotic normality of the maximum likelihood and variational estimators of the latent block model were finally tackled and obtained in a joint work with V. Brault and M. Mariadassou.

J-M. Poggi (with R. Genuer, C. Tuleau-Malot, N. Villa-Vialaneix), have published an article on random forests in “big data” classification problems, and have performed a review of available proposals about random forests in parallel environments as well as on online random forests. Three variants involving subsampling, Big Data-bootstrap and MapReduce respectively were tested on two massive datasets, one simulated one, and the other, real-world data.

With G. Lecué, Matthieu Lerasle worked on robust machine learning by median-of-means, providing an alternative to the Lugosi and Mendelson approach based on median of means for learning. This alternative is easier to present and to analyse theoretically. Furthermore, they proposed an algorithm to approximate this estimator, which could not be done for Lugosi and Mendelson’s champions of tournaments (submitted).

6.4. Estimation for conditional densities in high dimension

Participants: Claire Lacour, Jeanne Nguyen.

Jeanne Nguyen is working on estimation for conditional densities in high dimension. Much more informative than the regression function, conditional densities are of high interest in recent methods, particularly in the Bayesian framework (studying the posterior distribution). Considering a specific family of kernel estimators, she is studying a greedy algorithm for selecting the bandwidth. Her method addresses several issues: avoiding the curse of high dimensionality under some suitably defined sparsity conditions, being computationally efficient using iterative procedures, and early variable selection, providing theoretical guarantees on the minimax risk.

6.5. Reliability

Participants: Gilles Celeux, Florence Ducros, Patrick Pamphile.

Since June 2015, in the framework of a CIFRE convention with Nexter, Florence Ducros has begun 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. Statistical analysis of genomic data

Participants: Gilles Celeux, Christine Keribin, Yann Vasseur, Kevin Bleakley.

The subject of Yann Vasseur's PhD Thesis, supervised by Gilles Celeux and Marie-Laure Martin-Magniette (INRA URGV), was the inference of a regulatory network for Transcriptions Factors (TFs), which are specific genes, of *Arabidopsis thaliana*. For this, a transcriptome dataset with a similar number of TFs and statistical units was available. They reduced the dimension of the network to avoid high-dimensional difficulties. Representing this network with a Gaussian graphical model, the following procedure was defined:

1. *Selection step*: choose the set of TF regulators (supports) of each TF.
2. *Classification step*: deduce co-factor groups (TFs with similar expression levels) from these supports.

Thus, the reduced network would be built on the co-factor 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 into a number of co-factor groups, selected with BIC or the exact ICL criterion. Since these models are built in an ad hoc way, Yann Vasseur has defined complex simulation tools to asses their performances in a proper way.

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 has now been submitted.

In collaboration with researchers from the Pasteur Institute, Kevin Bleakley worked on statistical tests in the context of research into what leads to dengue fever *without symptoms* as opposed to *with* symptoms. This work was published in *Science Translational Medicine*.

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 is submitted.

6.7. 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 worked on the detection of associations between drugs and adverse events in the framework of the PhD of Valerie Robert, which was defended this year. At first, this team developed model-based clustering inspired by latent block models (LBMs), which consists of co-clustering rows and columns of two binary tables, imposing the same row ranking. This enabled 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 identifiability of the model, and some results are shown for simulated data. The exact ICL criterion has been extended to this double block latent model. Through computer experiments, Valérie Robert demonstrated the interest of the proposed model, compared with standard contingency table analysis, to detect co-prescription and masking effects.

Futhermore, with V. Robert, C. Kerebin and G. Celeux showed that it can be useful to use an LBM model on a contingency table of drugs and adverse effects to do cluster initialization for dealing with individual's data.

6.8. Statistical rating and ranking of scientific journals

Participants: Gilles Celeux, Julie Josse, Jean-Louis Foulley.

In collaboration with Jean-Louis Foulley (Montpellier University), Gilles Celeux and Julie Josse have done research on the statistical rating and ranking of scientific journals. They have proposed Dirichlet multinomial Bayesian models for pagerank-type algorithms allowing self-citations to be excluded. The resulting methods were tested on a set of 47 scientific journals.

6.9. Statistical mathematics

Participant: Matthieu Lerasle.

In collaboration with R. Diel, Matthieu Lerasle published an article on nonparametric estimation for random walks in random environments. They proposed a non-parametric approach for estimating the distribution of the environment from the observation of one trajectory of a random walk in it. They obtained risk bounds in sup-norm for the cumulative distribution function of the environment.

6.10. Random graph theory

Participant: Matthieu Lerasle.

In collaboration with R. Chetrite and R. Diel, Matthieu Lerasle published an article on the number of potential winners in the Bradley-Terry model in random environments. They proposed the first mathematical study of the Bradley-Terry model where the values of players are i.i.d. realisations of some distribution. They proved that a Bradley-Terry tournament is fair (in the sense that the best player ends up with the largest number of victories) under a certain convexity condition on the tail distribution of the values. They also showed that this condition is sharp and provided sharp estimate of the number of potential winners when the condition fails.

He also collaborated with R. Diel and S. Le Corff on learning latent structures of large random graphs, investigating the possibility of estimating latent structure in sparsely observed random graphs. The main example was a Bradley-Terry tournament where each team has only played a few games. It is well known that individual values of the teams cannot be consistently estimated in this setting. They showed that their distribution on the other hand can be, and provide general tools for bounding the risk of the maximum likelihood estimator (submitted).

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.

7.2. Bilateral Grants with Industry

Benjamin Auder and Jean-Michel Poggi are participants in the grant PGMO-IRSDI, in the *Research Initiative In Industrial Data Science* context, on the subject: Disaggregated Electricity Forecasting using Clustering of Individual Consumers.

8. Partnerships and Cooperations

8.1. Regional Initiatives

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

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.

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 Perugia, Italy

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 research, from late 2016 until early 2017.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

Sylvain Arlot organized (with Guillaume Charpiat) the Workshop Statistics/Learning at Paris-Saclay (2nd edition), at IHES (Bures-sur-Yvette).

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 Perugia, Italy.
- Sylvain Arlot is one of the co-organizers of the Junior Conference on Data Science and Engineering at Paris-Saclay (2nd edition in 2017).
- Jean-Michel Poggi was president of the Scientific Program Committee, ENBIS 2017, Naples, 10-14 June 2017.
- Jean-Michel Poggi was member of the Conference Scientific Board of IES 2017, Naples, Italy, 6-8 September 2017.

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*.

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 2016.

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.

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: Valérie Robert, 2013, Gilles Celeux and Christine Keribin. Defended in June 2017

PhD : Yann Vasseur, 2013, Gilles Celeux and Marie-Laure Martin-Magniette (URGV). Defended in December 2017

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 Brécheteau, 2015, Pascal Massart

PhD in progress: Hedi Hadiji, 2017, Pascal Massart

PhD in progress: Eddie Aamari, 2015, Pascal Massart and Frédéric Chazal

PhD: Damien Garreau, 2013, Sylvain Arlot and Gérard Biau (UPMPC). Defended in October 2017

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

Masters internship: Thomas Prochwicz. Christine Keribin conducted a preliminary study on expert aggregation by supervising this three month internship.

9.2.3. Juries

S. Arlot was a member of the Ph.D. jury of Jilai Mei (Université Paris-Sud).

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] E. AAMARI. *Convergence Rates for Geometric Inference*, Université Paris-Saclay, September 2017, https://hal.inria.fr/tel-01607782

Articles in International Peer-Reviewed Journals

[2] C. LACOUR, P. MASSART, V. RIVOIRARD. *Estimator selection: a new method with applications to kernel density estimation*, in "Sankhya A", August 2017, vol. 79, n^{o} 2, pp. 298 - 335, https://arxiv.org/abs/1607.05091 [*DOI :* 10.1007/s13171-017-0107-5], https://hal.archives-ouvertes.fr/hal-01346081

[3] E. SIMON-LORIERE, V. DUONG, A. TAWFIK, S. UNG, S. LY, I. CASADEMONT, M. PROT, N. COURTEJOIE, K. BLEAKLEY, P. BUCHY, A. TARANTOLA, P. DUSSART, T. CANTAERT, A. SAKUNTABHAI. *Increased adaptive immune responses and proper feedback regulation protect against clinical dengue*, in "Science Translational Medicine", August 2017, vol. 9, n^{o} 405, eaal5088 p. , https://arxiv.org/abs/1712.05692 [*DOI :* 10.1126/SCITRANSLMED.AAL5088], https://hal.inria.fr/hal-01656594

Invited Conferences

[4] J.-P. BAUDRY, G. CELEUX. *Assessing model-based clustering methods with cytometry data sets*, in "IFCS 2017 - Conference of the International Federation of Classification Societies", Tokyo, Japan, August 2017, https://hal.inria.fr/hal-01649085

International Conferences with Proceedings

[5] C. KERIBIN, G. CELEUX, V. ROBERT. *The Latent Block Model: a useful model for high dimensional data*, in "ISI 2017 - 61st world statistics congress", Marrakech, Morocco, July 2017, pp. 1-6, https://hal.inria.fr/hal-01658589

Conferences without Proceedings

[6] V. BRAULT, A. CHANNAROND, V. ROBERT. *Généralisation de l'algorithme Largest Gaps pour le modèle des blocs latents non-paramétrique*, in "49èmes Journées de Statistique", Avignon, France, May 2017, https://hal.archives-ouvertes.fr/hal-01510984

[7] V. BRAULT, C. KERIBIN, M. MARIADASSOU. *Équivalence asymptotique des vraisemblances observée et complète dans le modèle de blocs latents*, in "XXIV èmes Rencontres de la Société Francophone de Classification", Lyon, France, Société Francophone de Classification, June 2017, https://hal.archives-ouvertes.fr/hal-01510994

Other Publications

[8] E. AAMARI, J. KIM, F. CHAZAL, B. MICHEL, A. RINALDO, L. WASSERMAN. *Estimating the Reach of a Manifold*, May 2017, https://arxiv.org/abs/1705.04565 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01521955

[9] E. AAMARI, C. LEVRARD. *Non-Asymptotic Rates for Manifold, Tangent Space, and Curvature Estimation*, April 2017, https://arxiv.org/abs/1705.00989 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01516032

[10] E. AAMARI, C. LEVRARD. *Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction*, November 2017, https://arxiv.org/abs/1512.02857 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01245479

[11] S. ARLOT. *Cross-validation*, March 2017, https://arxiv.org/abs/1703.03167 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01485508

[12] S. ARLOT. *Tutorial on statistical learning*, March 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01485506

[13] V. BRAULT, C. KERIBIN, M. MARIADASSOU. *Consistency and Asymptotic Normality of Latent Blocks Model Estimators*, April 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01511960

[14] N. EL HAOUIJ, J.-M. POGGI, R. E. GHOZI, S. SEVESTRE-GHALILA, M. JAÏDANE. *Random Forest-Based Approach for Physiological Functional Variable Selection: Towards Driver's Stress Level Classification*, January 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01426752

[15] J.-L. FOULLEY, G. CELEUX, J. JOSSE. *Empirical Bayes approaches to PageRank type algorithms for rating scientific journals*, June 2017, working paper or preprint, https://hal.inria.fr/hal-01535134

[16] G. MAILLARD, S. ARLOT, M. LERASLE. *Cross-validation improved by aggregation: Agghoo*, September 2017, https://arxiv.org/abs/1709.03702 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01585595

[17] V. ROBERT, Y. VASSEUR. *Comparing high dimensional partitions with the Coclustering Adjusted Rand Index*, May 2017, working paper or preprint, https://hal.inria.fr/hal-01524832