IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2016

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:

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

1. Members

Research Scientists

Kevin Bleakley [Inria, Researcher]
Gilles Celeux [Inria, Senior Researcher]
Julie Josse [Inria, Researcher]
Matthieu Lerasle [CNRS, Researcher]

Faculty Members

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

Engineers

Benjamin Auder [CNRS, Researcher]
Josselin Demont [Inria]
Jonas Renault [Inria]
Christian Poli [Inria]

PhD Students

Benjamin Goehry [Univ. Paris XI]
Neska Haouij [Univ. Paris XI and Univ. Tunis El Manar]
Valerie Robert [Univ. Paris XI]
Yann Vasseur [Univ. Paris XI]
Jeanne Nguyen [Univ. Paris XI]
Florence Ducros [Cifre: Univ. Paris XI and Nexter]
Claire Brécheteau [Univ. Paris XI]
Eddie Aamari [Univ. Paris XI]
Damien Garreau [Univ. Paris XI]
Guillaume Maillard [Univ. Paris XI]

Post-Doctoral Fellow

Kaniav Kamary [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 aging.

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. 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 is completing 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 is dealing with high-dimensional ill-posed problems where the number of variable is almost equal to the number of observations. He has designed heuristic tools using regularized regression methods to circumvent this difficulty.

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 aims to predict dengue severity using only low-dimensional clinical data obtained at hospital arrival. A third 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.

4.5. 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 is completing a thesis co-supervised by Gilles Celeux and Christine Keribin on this topic, which involves 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: Mixture model - Block cluster analysis
SCIENTIFIC DESCRIPTION

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 also been provided. This non-asymptotic criterion appears to be more relevant than the BIC-like approximation of ICL.
FUNCTIONAL DESCRIPTION

BlockCluster is an R package for co-clustering of binary, contingency and continuous data based on mixture models.

- Participants: Gilles Celeux, Christine Keribin, Christophe Biernacki and Serge Iovleff
- Contact: Gilles Celeux
- URL: http://cran.r-project.org/web/packages/blockcluster/index.html

5.2. Mixmod

Multi-purpose software for model-based clustering and classification with continuous and categorical variables.
KEYWORDS: Mixture model - cluster analysis - discriminant analysis
FUNCTIONAL DESCRIPTION

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.

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

Jonas Renault, an engineer, is in charge of developing a web version of MIXMOD.

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

5.3. 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 developed by MODAL team (Inria Lille). BlockCluster is a library for co-clustering 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.

- Participants: Christophe Biernacki, Gilles Celeux, Benjamin Auder, Josselein Demont, Jonas Renault
- Contact: Jonas Renault
- URL: https://massiccc.lille.inria.fr

6. New Results

6.1. Model selection in Regression and Classification

Participants: Gilles Celeux, Serge Cohen, 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 proposed to sort 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. 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. This year their method has been compared favorably to competing methods.

Sylvain Arlot, in collaboration with Damien Garreau (Inria Paris, Sierra team), studied the kernel change-point algorithm (KCP) proposed by Arlot, Celisse and Harchaoui, that 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. They 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 minimax rate $\log(n)/n$. 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 $\mathbb{R}^d$). 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.

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. 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 time series forecasting models adapted to each cluster.

J-M. Poggi, with J. Cugliari, Y. Goude, have proposed building clustering tools useful for forecasting load consumption. The idea is to disaggregate the global signal in such a way that the sum of disaggregated forecasts significantly improves the prediction of the whole global signal. The strategy has three steps: first they cluster curves defining super-consumers, then they build a hierarchy of partitions from which the best one is selected with respect to a disaggregated forecast criterion. The proposed strategy is applied to a dataset of individual consumers from the French electricity provider EDF.

V. Thouvenot and J-M. Poggi, with A. Pichavant, A. Antoniadis, Y. Goude, consider electricity forecasting using multi-stage estimators of nonlinear additive models. An automatic procedure for variable selection is used to correct middle term forecasting errors for short term forecasting. An application to the EDF customer load demand at an aggregate level is considered as well as an application on load demand from the GEFCom 2012 competition; this is a local application.

6.2. Estimator selection

Participants: Claire Lacour, Pascal Massart.

Estimator selection has become a crucial issue in nonparametric estimation. Two widely used methods are penalized empirical risk minimization (such as penalized log-likelihood estimation) and pairwise comparison (such as Lepski's method). C. Lacour, P. Massart and V. Rivoirard have developed a new method for bandwidth selection which is in some sense intermediate between these two main methods mentioned above, and is called "Penalized Comparison to Overfitting". They have first provided some theoretical results (oracle bounds, minimal penalty) within the framework of kernel density estimation, which leads to some fully data-driven selection strategies. Currently, S. Varet is implementing this method, making a thorough comparison with other selection methods, and tackling the multivariate case. Theoretical work is also in progress, in order to expand the method to other loss functions, such as the Hellinger loss.

6.3. Statistical learning methodology and theory

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

Gaussian graphical models are widely used to infer and visualize networks of dependencies between continuous variables. However, inferring the graph is difficult when the sample size is small compared to the number of variables. To reduce the number of parameters to estimate in the model, the past PhD. students Emilie Devijver (supervisors: Pascal Massart and Jean-Michel Poggi) and Mélina Gallopin (supervisor: Gilles Celeux) proposed a non-asymptotic model selection procedure supported by strong theoretical guarantees based on an oracle inequality and a minimax lower bound. The covariance matrix of the model is approximated by a block-diagonal matrix. The structure of this matrix is detected by thresholding the sample covariance matrix, where the threshold is selected using the slope heuristic. Based on the block-diagonal structure of the covariance matrix, the estimation problem is divided into several independent problems: subsequently, the network of dependencies between variables is inferred using the graphical lasso algorithm in each block. The performance of the procedure has been illustrated on simulated data. An application to a real gene expression dataset with a limited sample size has been achieved: the dimension reduction allows attention to be objectively focused on interactions among smaller subsets of genes, leading to a more parsimonious and interpretable modular network. This work has been accepted for publication in the *Journal of the American Statistical Association*.

J-M. Poggi, with A. Bar-Hen, have focused on individual observation diagnosis issues for graphical models. The use of an influence measure is a classical diagnostic method to measure the perturbation induced by single elements. The stability issue is here considered using jackknife. For a given graphical model, tools to perform diagnosis on observations are provided. In the second step, a filtering of the dataset to obtain a stable network is proposed.

Latent Block Models (LBM) are a model-based method to cluster simultaneously the d columns and n rows of a data matrix. The Blockcluster package estimates such LBMs. Parameter estimation in LBM is a difficult and multifaceted problem. Although various estimation strategies have been proposed and are now well-understood empirically, theoretical guarantees about their asymptotic behavior is rather rare. Christine Keribin, in collaboration with Mahendra Mariadassou (INRA) and Vincent Brault (Université de Grenoble) have shown that under some mild conditions on the parameter space, and in an asymptotic regime where $\log(d)/n$ and $\log(n)/d$ go to 0 when n and d go to $+\infty$, (1) the maximum likelihood estimate of the complete model (with known labels) is consistent and (2) the log-likelihood ratios are equivalent under the complete and observed (with unknown labels) models. This equivalence allows us to transfer the asymptotic consistency to the maximum likelihood estimate under the observed model. Moreover, the variational estimator is also consistent. These results extends the results of Bickel et al. (2013) on stochastic block models, and detail the case where the parameter exhibits symmetry.

For the same LBM, Valérie Robert and Yann Vasseur have extended the popular Adjusted Rand Index (ARI) to the task of simultaneous clustering of the rows and columns of a given matrix. This new index, called the Coclustering Adjusted Rand Index (CARI), overcomes the label switching phenomenon while remaining useful and competitive with respect to other indices. Indeed, partitions with high numbers of clusters can be considered, and no convention is required when the numbers of clusters in partitions are different. They are now exploring links with other indices.

Gilles Celeux continued his collaboration with Jean-Patrick Baudry on model-based clustering. This year, they proposed to consider the model selection criterion ICL as a validity index. They show how it can be coupled with a null model of homogeneity focusing on clustering. This null model, which includes the Gaussian distributions, can be difficult to analyze. They find an explicit representation for simple models and show how the parametric bootstrap test can be applied in such situations. In more general situations, they propose a solution for applying this approach involving an "acceptance-rejection" procedure which explores the parameter space to approximate the maximum likelihood estimator inside the null model of homogeneity. The uncovering of this null model highlights the notion of class underlying ICL, and confirms the results of earlier results which show that ICL is consistent for a loss function taking clustering into account.

In collaboration with Arthur White and Jason Wyse (Trinity College, Dublin) Gilles Celeux has evaluated for multivariate Poisson mixture models the performance of a greedy search method compared to the expectation maximization (EM) algorithm, to optimize the ICL model selection criterion, which can be computed exactly

for such models. It appears that EM gives often slightly better results, but the greedy search is computationally is more efficient.

The Dutch and French schools of data analysis differ in their approaches to the question: How does one understand and summarize the information contained in a data set? Julie Josse, in collaboration with François Husson (Agro Rennes) and Gibert Saporta (CNAM, Paris), explored the shared factors and differences between the schools, with a focus on methods dedicated to the analysis of categorical data, which are known either as homogeneity analysis (HOMALS) or multiple correspondence analysis (MCA). MCA is a dimension-reduction method which plays a large role in the analysis of tables with categorical nominal variables such as survey data. Though it is usually motivated and derived using geometric considerations, they proved that it amounts to a single proximal Newton step of a natural bilinear exponential family model for categorical data: the multinomial logit bilinear model. They compared and contrasted the behavior of MCA with that of the model on simulations, and discussed new insights into the properties of both exploratory multivariate methods and their cognate models. The main conclusion is to recommend approximating the multilogit model parameters using MCA. Indeed, estimating the parameters of the model is not a trivial task, whereas MCA has the great advantage of being easily solved by a singular value decomposition, as well as being scalable to large datasets.

Julie Josse, with Sobczyk and Bogdan, have discussed the problem of estimating the number of principal components in Principal Components Analysis (PCA). They address this issue by presenting an approximate Bayesian approach based on Laplace approximation, and introduce a general method for building the model selection criteria, called PEnalized SEmi-integrated Likelihood (PESEL). This general framework encompasses a variety of existing approaches based on probabilistic models, like e.g., Bayesian Information Criterion for the Probabilistic PCA (PPCA), and allows for construction of new criteria, depending on the size of the data set at hand. Specifically, they define PESEL when the number of variables substantially exceeds the number of observations. Numerical simulations show that PESEL-based criteria can be quite robust against deviations from probabilistic model assumptions. Selected PESEL-based criteria for estimation of the number of principal components are implemented in the R package varclust, which is available on Github.

Gillies Celeux and Julie Josse have started research on missing data for model-based clustering in collaboration with Christophe Biernacki (Modal, Inria Lille). The aim of this research is to propose appropriate and efficient tools for the packages Mixmod and Mixtcomp.

In collaboration with Jean-Michel Marin (Université de Montpellier) and Christian Robert (Université Paris 9-Dauphine), Gilles Celeux and Kaniav Kamary investigated the ability of Bayesian inference to properly estimate the parameters of Gaussian mixtures in high dimensions. Their study shows how the choice of the prior distributions is important. In particular, independent prior distributions give much better performances. Moreover, when the dimension d becomes very large (say $d > 40$) Bayesian inference becomes questionable. The results of this study will be gathered in a chapter of a book on mixture models that Gilles Celeux is preparing with Christian Robert and Sylvia Fruhwirth Schnatter.

Sylvain Arlot, in collaboration with Robin Genuer (ISPED), studied the reasons why random forests work so well in practice. Focusing on the problem of quantifying the impact of each ingredient of random forests on their performance, they showed that such a quantification is possible for a simple pure forest, leading to conclusions that could apply more generally. Then, they considered "hold-out" random forests, which are a good midpoint between "toy" pure forests and Breiman's original random forests.

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 stress levels during a 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 for 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 expert-based studies of physiological signals.

J-M. Poggi (with R. Genuer, C. Tuleau-Malot, N. Villa-Vialaneix), have focused 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 are tested on two massive datasets, one simulated one, and the other, real-world data.

B. Auder and J-M. Poggi (with M. Bobbia, B. Portier) have tested some methods for sequential aggregation for forecasting PM10 concentrations for the next day, in the context of air quality monitoring in Normandy (France). The main originality is that the set of experts contains at the same time statistical models built by means of various methods and groups of predictors, as well as experts coming from deterministic chemical models of prediction. The obtained results show that such a strategy clearly improves the performances of the best expert both in terms of prediction errors and in terms of alerts. What is more, it obtains, for the non-convex weighting strategy, the "unbiasedness" of observed-forecasted scatterplots, which is extremely difficult to obtain.

J-M. Poggi (with A. Antoniadis, I. Gijbels, S. Lambert-Lacroix) have considered the joint estimation and variable selection for mean and dispersion in proper dispersion models. They used recent results on Bregman divergence for establishing theoretical results for the proposed estimators in fairly general settings, and also studied variable selection when there is a large number of covariates, with this number possibly tending to infinity with the sample size. The proposed estimation and selection procedure is investigated via a simulation study, and illustrated via some real data applications.

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. It involves the estimation of mixtures and competing risk models in a highly-censored setting. Moreover, she can deduce from these models operational tools to estimate the number of spare parts to be stocked in a given period. These tools are defined to take vehicle use patterns into account.

6.6. Statistical analysis of genomic data

Participants: Gilles Celeux, Mélina Gallopin, 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), is 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 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-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 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. Moreover, the performance of `HTSCluster` is compared with $k$means-like algorithms using the χ^2 distance.

In collaboration with Benno Schwikowski, Iryna Nikolayeva and A Anavaj Sakuntabhai (Pasteur Institute, Paris), Kevin Bleakley works 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.

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 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 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 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 has demonstrated the interest of the proposed model, compared with standard contingency table analysis, to detect co-prescription and masking effects.

6.8. Statistical rating and ranking of scientific journals

Participants: Gilles Celeux, Julie Josse, Simon Grah.

In collaboration with Jean-Louis Foulley (université of Montpellier), Gilles Celeux and Julie Josse have started research on statistical rating and ranking of scientific journals. This research was the subject of the internship of Simon Grah (Université Paris-Sud). Simon Grah compared many models on a set of 47 statistical journals. His study showed that the Row-Column (RC) models appears to be the most relevant. In the future, Bayesian inference for different approaches, including PageRank, will be considered.

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.

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.

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

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 Paris.

Julie Josse was chair of useR!2016, Stanford, CA, USA, July 2016. http://user2016.org/

Juile Josse is member of the R foundation.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

Sylvain Arlot:

- Sylvain Arlot organized a workshop at IHES on statistics and learning in the Paris-Saclay area.
- Sylvain Arlot co-organized (with Francis Bach, Inria Paris, and Alain Celisse, Univ. Lille 1) a 2-day workshop at IHES about computational and statistical trade-offs in learning.

Jean-Michel Poggi:

- Organization Special Invited Session entitled “Advances in Random Forests”, 22nd International Conference on Computational Statistics, Oviedo, Spain, 23-26 August 2016.

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 Paris.

Sylvain Arlot co-organized the 1st Junior Conference on Data Science and Engineering at Paris-Saclay (at LAL, Orsay).

9.1.2. Scientific Events Selection

9.1.2.1. Chair of Conference Program Committees

Jean-Michel Poggi was:

- President of the Scientific Programme Committee, ENBIS 2017, Naples, 10-14 June 2017

9.1.2.2. Member of the Conference Program Committees

Jean-Michel Poggi was:

- Member of the Scientific Committee of CESS 2016, Conference of European Statistics Stakeholders, Hungarian Academy of Sciences, Budapest, 20-21 October 2016
- Member of SPC ENBIS-2016, Sheffield, UK, 11-15 September 2016
- Member of the Scientific committee of the journées MAS 2016, Grenoble
- Member of SPC COMPSTAT 2016, 22nd International Conference on Computational Statistics, Oviedo, Spain, 23-26 August 2016.

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*.

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 2016.

9.1.5. 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)

9.1.6. Scientific Expertise

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

9.1.7. 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.

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 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 Brécheteau, 2015, Pascal Massart

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

PhD in progress: Damien Garreau, 2016, Sylvain Arlot and Gérard Biau

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

PhD in progress: Jeanne Nguyen, 2015, Claire Lacour

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

9.2.3. Juries

- Ph.D. Jérémy Bensadon: Sylvain Arlot (president)
- Ph.D. Gwenaelle Mabon: Sylvain Arlot (member)
- Ph.D. Mokhtar Alaya: Sylvain Arlot (president)
- Ph.D. Marie-Liesse Cauwet: Sylvain Arlot (member)

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] S. ARLOT, R. GENUER. *Comments on: "A Random Forest Guided Tour" by G. Biau and E. Scornet*, in "Test", 2016, vol. 25, n° 2, pp. 228–238, The final publication is available at Springer: http://dx.doi.org/10.1007/s11749-016-0484-4 [*DOI :* 10.1007/s11749-016-0484-4], https://hal.archives-ouvertes.fr/hal-01297557

[2] G. BIAU, K. BLEAKLEY, B. CADRE. *The Statistical Performance of Collaborative Inference*, in "Journal of Machine Learning Research", 2016, https://hal.archives-ouvertes.fr/hal-01170254

[3] F. CHAZAL, P. MASSART, B. MICHEL. *Rates of convergence for robust geometric inference*, in "Electronic journal of statistics ", 2016, vol. 10, n° 2, 44 p. , https://hal.inria.fr/hal-01336913

[4] C. LACOUR, P. MASSART. *Minimal penalty for Goldenshluger-Lepski method*, in "Stochastic Processes and their Applications", December 2016, vol. 126, n° 12, pp. 3774–3789 [*DOI :* 10.1016/J.SPA.2016.04.015], https://hal.archives-ouvertes.fr/hal-01121989

International Conferences with Proceedings

[5] M. MARIADASSOU, V. BRAULT, C. KERIBIN. *Normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle de blocs latents*, in "48èmes Journées de Statistique de la SFdS", Montpellier, France, 48èmes Journées de Statistique de la SFdS, May 2016, https://hal.inria.fr/hal-01440084

Conferences without Proceedings

[6] J.-P. BAUDRY, G. CELEUX. *What does ICL tell us about homogeneity for Model-Based Clustering?*, in "Workshop on Model-Based Clustering and Classification", Catane, Italy, Salvatore Ingrassia, September 2016, https://hal.inria.fr/hal-01437468

[7] G. CELEUX, A. WHITE, J. WYSE. *Comparing EM with a greedy search algorithm to optimize ICL for Poisson mixture models* , in "Worshop on Model-Based Clustering", Paris, France, July 2016, https://hal.inria.fr/hal-01437459

Other Publications

[8] E. AAMARI, C. LEVRARD. *Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction*, June 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01245479

[9] S. ARLOT, A. CELISSE, Z. HARCHAOUI. *A kernel multiple change-point algorithm via model selection*, March 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-00671174

[10] 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

[11] D. GARREAU, S. ARLOT. *Consistent change-point detection with kernels*, December 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01416704

[12] C. LACOUR, P. MASSART, V. RIVOIRARD. *Estimator selection: a new method with applications to kernel density estimation*, July 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01346081

[13] P. SOBCZYK, M. BOGDAN, J. JOSSE. *Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood*, 2016, 31 pages, 7 figures, https://hal.archives-ouvertes.fr/hal-01342815