IN PARTNERSHIP WITH:
CNRS

Université Lille 2

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2017

Project-Team MODAL

MOdel for Data Analysis and Learning

IN COLLABORATION WITH: Laboratoire Paul Painlevé (LPP)

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team MODAL

Creation of the Team: 2010 September 01, updated into Project-Team: 2012 January 01

Keywords:

Computer Science and Digital Science:

A3.1.4. - Uncertain data
A3.2.3. - Inference
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.5. - Bayesian methods
A3.4.7. - Kernel methods
A5.2. - Data visualization
A6.2.3. - Probabilistic methods
A6.2.4. - Statistical methods
A6.3.3. - Data processing
A9.2. - Machine learning

Other Research Topics and Application Domains:

B1.1.6. - Genomics
B2.2.3. - Cancer
B9.4.5. - Data science
B9.5.3. - Economy, Finance
B9.5.5. - Sociology

1. Personnel

Research Scientists

Pascal Germain [Inria, Researcher, from Nov 2017]
Benjamin Guedj [Inria, Researcher]

Faculty Members

Christophe Biernacki [Team leader, Univ des sciences et technologies de Lille, Professor, HDR]
Alain Celisse [Univ des sciences et technologies de Lille, Associate Professor]
Serge Iovleff [Univ des sciences et technologies de Lille, Associate Professor]
Guillemette Marot [Univ du droit et de la santé de Lille, Associate Professor]
Cristian Preda [Univ des sciences et technologies de Lille, Professor, HDR]
Vincent Vandewalle [Univ du droit et de la santé de Lille, Associate Professor]

Post-Doctoral Fellow

Alexandru Amarioarei [Inria, until Feb 2017]

PhD Students

Yaroslav Averyanov [Inria, from Sep 2017]
Maxime Baelde [A-Volute]
Anne Lise Bedenel [MeilleureAssurance]
Maxime Brunin [Univ des sciences et technologies de Lille, until Sep 2017]

Adrien Ehrhardt [CA CF]
Le Li [Univ. d'Angers & iAdize]

Technical staff

Matthieu Marbac Lourdelle [Inria, until Aug 2017]
Bhargav Srinivasa Desikan [Inria, from Oct 2017]

Interns

Miguel Assuncao [Inria, until Feb 2017]
Mohamed Biaz [Inria, from Jun 2017 until Sep 2017]
Julien Gheysens [Inria, from May 2017 until Aug 2017]
Thierry Mottet [from Apr 2017 until Aug 2017]
Nicolas Pompidor [Inria, from Jun 2017 until Sep 2017]
Matthieu Rousseaux [from Mar 2017 until Jul 2017]
Bhargav Srinivasa Desikan [Inria, until Jul 2017]

Administrative Assistant

Anne Rejl [Inria]

Visiting Scientists

Mohamed Aimen Ben Hajkacem [Institut Supérieur de Gestion (Tunis), Mar 2017]
Maxime Brunin [Univ du droit et de la santé de Lille, from Oct 2017]

External Collaborators

Faicel Chamroukhi [Univ de Caen Basse-Normandie]
Sophie Dabo [Univ Charles de Gaulle]
Julien Jacques [Univ Lumière Lyon, HDR]
Philippe Heinrich [Univ des sciences et technologies de Lille]

2. Overall Objectives

2.1. Overall Objectives

Modal is a team focused on statistical methodology for data analysis (clustering, visualization) and learning (classification, density estimation, aggregation, matrix factorization, ...). In this context, the core of the team's work is to design meaningful generative models for prominent complex data (mixed structured data), which are still almost ignored in the literature. From those generative models, learning procedures are proposed.

The scientific objectives of Modal include the two following methodological directions: generative model design and data visualization through such models. In each case, several means of dissemination are considered towards academic and/or industrial communities: publications in international journals (in statistics or biostatistics), workshops to raise or identify emerging topics, and publicly available specific software relying on the proposed new methodologies.

3. Research Program

3.1. Generative model design

The first objective of MODAL consists in designing, analyzing, estimating and evaluating new generative parametric models for multivariate and/or heterogeneous data. It corresponds typically to continuous and categorical data but it includes also other widespread ones like ordinal, functional, ranks,...Designed models have to take into account potential correlations between variables while being (1) justifiable and realistic, (2) meaningful and parsimoniously parameterized, (3) of low computational complexity. The main purpose is to identify a few theoretical and general principles for model generation, loosely dependent on the variable nature. In this context, we propose two concurrent approaches which could be general enough for dealing with correlation between many types of homogeneous or heterogeneous variables:

- Designs general models by combining two extreme models (full dependent and full independent) which are well-defined for most of variables;
- Uses kernels as a general way for dealing with multivariate and heterogeneous variables.

3.2. Data visualization

The second objective of MODAL is to propose meaningful and quite accurate low dimensional visualizations of data typically in two-dimensional (2D) spaces, less frequently in one-dimensional (1D) or three-dimensional (3D) spaces, by using the generative models designed in the first objective. We propose also to visualize simultaneously the data and the model. All visualizations will depend on the aim at hand (typically clustering, classification or density estimation). The main originality of this objective lies in the use of models for visualization, a strategy from which we expect to have a better control on the subjectivity necessarily induced by any graphical display. In addition, the proposed approach has to be general enough to be independent on the variable nature. Note that the visualization objective is consistent with the dissemination of our methodologies through specific softwares. Indeed, displaying data is an important step in the data analysis process.

4. Application Domains

4.1. Economic world

The Modal team applies it research to the economic world through CIFRE Phd supervision such as CA CF (credit scoring), A-Volute (expert in 3D sound), Meilleur Taux (insurance comparator), ... It also has many contracts with companies such as Running care (sport and medical coaching), Arcelor-Mittal (steel industry), Alstom (integrated transport systems) and Vallourec (tubular solutions for the energy markets and for other demanding industrial applications).

4.2. Biology

The second main application domain of the team is the biology. Members of the team are involved in the supervision and scientific animation of the bilille platform, the bioinformatics and bioanalysis platform of Lille. A 2 months research contact has been performed for the Florimond Desprez company (seed industry), and several academic projects have been led in collaboration with research teams in health and biology.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. NIPS 2017 workshop

Benjamin Guedj, Pascal Germain (both at Modal) and Francis Bach (SIERRA, Inria Paris) co-organize a NIPS 2017 workshop, called “(Almost) 50 shades of Bayesian learning: PAC-Bayesian trends and insights". A large audience is expected, and the workshop has a series of prestigious international speakers. See the website.

5.1.2. Recruitment of a new researcher

Pascal Germain has been recruited has CR2 in the team, three years after the recruitment of Benjamin Guedj the first CR recruited in the team.

6. New Software and Platforms

6.1. MixtComp

Mixture Computation

KEYWORDS: Clustering - Statistics - Missing data
FUNCTIONAL DESCRIPTION: MixtComp (Mixture Computation) is a model-based clustering package for mixed data originating from the Modal team (Inria Lille). It has been engineered around the idea of easy and quick integration of all new univariate models, under the conditional independence assumption. New models will eventually be available from researches, carried out by the Modal team or by other teams. Currently, central architecture of MixtComp is built and functionality has been field-tested through industry partnerships. Three basic models (Gaussian, multinomial, Poisson) are implemented, as well as two advanced models (Ordinal and Rank). MixtComp has the ability to natively manage missing data (completely or by interval). MixtComp is used as an R package, but its internals are coded in C++ using state of the art libraries for faster computation.

- Participants: Christophe Biernacki, Étienne Goffinet, Matthieu Marbac-Lourdelle, Quentin Grimonprez, Serge Iovleff and Vincent Kubicki
- Contact: Christophe Biernacki
- URL: https://modal-research.lille.inria.fr/BigStat

6.2. 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

6.3. CloHe

Clustering of Mixed data
KEYWORDS: Classification - Clustering - Missing data
FUNCTIONAL DESCRIPTION: Software of classification for mixed data with missing values with application to multispectral satellite image time-series

- Partners: CNRS - INRA
- Contact: Serge Iovleff
- URL: https://modal.lille.inria.fr/CloHe/

6.4. PACBayesianNMF

KEYWORDS: Statistics - Machine learning

FUNCTIONAL DESCRIPTION: Implementing NMF with a PAC-Bayesian approach relying upon block gradient descent

- Participants: Benjamin Guedj and Astha Gupta
- Contact: Benjamin Guedj
- URL: https://github.com/astha736/PACbayesianNMF

6.5. pycobra

KEYWORDS: Statistics - Data visualization - Machine learning
SCIENTIFIC DESCRIPTION: pycobra is a python library for ensemble learning, which serves as a toolkit for regression, classification, and visualisation. It is scikit-learn compatible and fits into the existing scikit-learn ecosystem.

pycobra offers a python implementation of the COBRA algorithm introduced by Biau et al. (2016) for regression.

Another algorithm implemented is the EWA (Exponentially Weighted Aggregate) aggregation technique (among several other references, you can check the paper by Dalalyan and Tsybakov (2007).

Apart from these two regression aggregation algorithms, pycobra implements a version of COBRA for classification. This procedure has been introduced by Mojirsheibani (1999).

pycobra also offers various visualisation and diagnostic methods built on top of matplotlib which lets the user analyse and compare different regression machines with COBRA. The Visualisation class also lets you use some of the tools (such as Voronoi Tesselations) on other visualisation problems, such as clustering.

- Participants: Bhargav Srinivasa Desikan and Benjamin Guedj
- Contact: Benjamin Guedj
- URL: https://github.com/bhargavvader/pycobra

6.6. STK++

Statistical ToolKit
KEYWORDS: Statistics - Linear algebra - Framework - Learning - Statistical learning
FUNCTIONAL DESCRIPTION: STK++ (Statistical ToolKit in C++) is a versatile, fast, reliable and elegant collection of C++ classes for statistics, clustering, linear algebra, arrays (with an API Eigen-like), regression, dimension reduction, etc. The library is interfaced with lapack for many linear algebra usual methods. Some functionalities provided by the library are available in the R environment using rtkpp and rtkore.

STK++ is suitable for projects ranging from small one-off projects to complete data mining application suites.

- Participant: Serge Iovleff
- Contact: Serge Iovleff
- URL: http://www.stkpp.org

6.7. rtkore

STK++ core library integration to R using Rcpp
KEYWORDS: C++ - Data mining - Clustering - Statistics - Regression

FUNCTIONAL DESCRIPTION: STK++ (http://www.stkpp.org) is a collection of C++ classes for statistics, clustering, linear algebra, arrays (with an Eigen-like API), regression, dimension reduction, etc. The integration of the library to R is using Rcpp. The rtkore package includes the header files from the STK++ core library. All files contain only templated classes or inlined functions. STK++ is licensed under the GNU LGPL version 2 or later. rtkore (the stkpp integration into R) is licensed under the GNU GPL version 2 or later. See file LICENSE.note for details.

- Participant: Serge Iovleff
- Contact: Serge Iovleff
- URL: https://cran.r-project.org/web/packages/rtkore/index.html

6.8. Platforms

6.8.1. MASSICCC Platform

MASSICCC is a demonstration platform giving access through a SaaS (service as a software) concept to data analysis libraries developed at Inria. It allows to obtain results either directly through a website specific display (specific and interactive visual outputs) or through an R data object download. It started in October 2015 for two years and is common to the Modal team (Inria Lille) and the Select team (Inria Saclay). In 2016, two packages have been integrated: Mixmod and MixtComp (see the specific section about MixtComp). In 2017, the BlockCluster package has been integrated and also a particular attention to provide meaningful graphical outputs (for Mixmod, MixtComp and BlockCluster) directly in the web platform itself has led to some specific developments.

7. New Results

7.1. An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization

Participant: Benjamin Guedj.

The quasi-Bayesian perspective has been extended to the popular setting of non-negative matrix factorization. This is a pivotal problem in machine learning (image segmentation, recommendation systems, audio source separation, ...) and an original estimator of the unobserved matrix has been proposed. An oracle inequality is derived, along with several possible implementations. This work is published in Mathematical Methods of Statistics [12].

It a joint work with Pierre Alquier from ENSAE - Université Paris-Saclay.

7.2. Simpler PAC-Bayesian Bounds for Hostile Data

Participant: Benjamin Guedj.

An original and much simpler way of deriving PAC-Bayesian bounds has been introduced through the use of f-divergences (therefore generalizing earlier works on Renyi's divergence and Kullback-Leibler divergence). This work is published in Machine Learning [13].

It a joint work with Pierre Alquier from ENSAE - Université Paris-Saclay.

7.3. Highlight 1 High-dimensional Adaptive Ranking with PAC-Bayesian Bounds

Participant: Benjamin Guedj.

The quasi-Bayesian perspective has been extended to the popular setting high-dimensional ranking. This is a pivotal problem in machine learning and is at the core of several applications in industry (recommender systems, active learning, ...). An original estimator of the scoring function is proposed, and we have shown its minimax optimal properties. Our procedure is adaptive to the unknown sparsity level of the data. This work is published in Journal of Statistical Planning and Inference.

It a joint work with Sylvain Robbiano from University College London.

7.4. Online Adaptive Clustering

Participant: Benjamin Guedj.

The quasi-Bayesian perspective has been extended to online adaptive clustering. Data streams are clustered dynamically with a quasi-Bayesian-flavored predictor, and we have proven minimax regret bounds. An efficient MCMC-based implementation is proposed.

7.5. Study of Transcriptional Regulation

Participant: Guillemette Marot.

The implementation of a mixture model of normal and exponential laws enabled to define a threshold on the number of co-recruiting transcriptional regulators in order to classify cis-regulatory modules. The new findings in Biology have been published in [16].

7.6. Functional Binary Linear Models for Stratified Samples

Participant: Sophie Dabo-Niang.

Sophie Dabo-Niang's new result concern a work on functional binary linear models for stratified samples. This work introduces a new functional binary choice model in a case-control or choice-based sample design context, where the response is binary, while the explanatory variable is functional. The model is estimated when the sample is stratified with respect to the values of the response variable. A dimensional reduction of the space of the explanatory random function based on a Karhunen-Loève expansion is used to define a conditional maximum likelihood estimate of the model. Based on this formulation, several asymptotic properties are given. Numerical experiments are used to compare the proposed method with the ordinary maximum likelihood method, which ignores the nature of the sampling. The proposed model yields encouraging results.

7.7. Mixture Model for Mixed Kind of Data

Participants: Christophe Biernacki, Matthieu Marbac-Lourdelle, Vincent Vandewalle.

A mixture model of Gaussian copula allows to cluster mixed kind of data has been proposed. Each component is composed by classical margins while the conditional dependencies between the variables is modeled by a Gaussian copula. The parameter estimation is performed by a Gibbs sampler. This work has been now published to an international journal [18]. Furthermore, an R package (MixCluster) is available on Rforge.

7.8. Data Units Selection in Statistics

Participant: Christophe Biernacki.

Usually, the data unit definition is fixed by the practitioner but it can happen that he/her hesitates between several data unit options. In this context, it is highlighted that it is possible to embed data unit selection into a classical model selection principle. The problem is introduced in a regression context before to focus on the model-based clustering and co-clustering context, for data of different kinds (continuous, count, categorical). This work is now in revision for an international journal [36]. It has led also to three invitations as a plenary session speaker to international or national conferences (the US Classification Society Conference [27], the French Classification Society Conference [28], the StatLearn conference [20]).

It is a joint work with Alexandre Lourme from University of Bordeaux.

7.9. Trade-off Between Computation Time and Accuracy

Participants: Christophe Biernacki, Maxime Brunin, Alain Célisse.

Most estimates practically arise from algorithmic processes aiming at optimizing some standard, but usually only asymptotically relevant, criteria. Thus, the quality of the resulting estimate is a function of both the iteration number and also the involved sample size. An important question is to design accurate estimates while saving computation time, and we address it in the simplified context of linear regression here. Fixing the sample size, we focus on estimating an early stopping time of a gradient descent estimation process aiming at maximizing the likelihood. It appears that the accuracy gain of such a stopping time increases with the number of covariates, indicating potential interest of the method in real situations involving many covariates. This work has been presented to an international conference [20] and a national conference [29], and a preprint is still being in progress.

Maxime Brunin will defend his PhD thesis related to this topic on January 2018.

7.10. Projection Under Pairwise Control

Participant: Christophe Biernacki.

Visualization of high-dimensional and possibly complex (non continuous for instance) data onto a low-dimensional space may be difficult. Several projection methods have been already proposed for displaying such high-dimensional structures on a lower-dimensional space, but the information lost is not always easy to use. Here, a new projection paradigm is presented to describe a non-linear projection method that takes into account the projection quality of each projected point in the reduced space, this quality being directly available in the same scale as this reduced space. More specifically, this novel method allows a straightforward visualization data in $\mathbb{R}^2$ with a simple reading of the approximation quality, and provides then a novel variant of dimensionality reduction.

This work is still under revision in an international journal [39].

It is a joint work with Hiba Alawieh and Nicolas Wicker, both from University of Lille.

7.11. Real-time Audio Sources Classification

Participants: Christophe Biernacki, Maxime Baelde.

Recent research on machine learning focuses on audio source identification in complex environments. They rely on extracting features from audio signals and use machine learning techniques to model the sound classes. However, such techniques are often not optimized for a real-time implementation and in multi-source conditions. It is proposed here a new real-time audio single-source classification method based on a dictionary of sound models (that can be extended to a multi-source setting). The sound spectrums are modeled with mixture models and form a dictionary. The classification is based on a comparison with all the elements of the dictionary by computing likelihoods and the best match is used as a result. It is found that this technique outperforms classic methods within a temporal horizon of 0.5s per decision (achieved 6errors on a database composed of 50 classes). This work has been now extended with success to the multi-sources classification case and also the computational load has been sufficiently reduced to reach the real time target (less than 50ms). This work has been presented to an international conference in Signal Processing [25] and also to a national conference [26]. A preprint is well advanced and should be submitted to an international journal at the end of 2017.

It is a joint work with Raphaël Greff, from the A-Volute company.

7.12. Model-Based Co-clustering for Ordinal Data

Participant: Christophe Biernacki.

A model-based co-clustering algorithm for ordinal data is presented. This algorithm relies on the latent block model embedding a probability distribution specific to ordinal data (the so-called BOS or Binary Ordinal Search distribution). Model inference relies on a Stochastic EM algorithm coupled with a Gibbs sampler, and the ICL-BIC criterion is used for selecting the number of co-clusters (or blocks). The main advantage of this ordinal dedicated co-clustering model is its parsimony, the interpretability of the co-cluster parameters (mode, precision) and the possibility to take into account missing data. Numerical experiments on simulated data show the efficiency of the inference strategy, and real data analyses illustrate the interest of the proposed procedure. The resulting work is in revision to an international journal [40].

This is joint work Julien Jacques from University of Lyon 2.

7.13. Model-Based Co-clustering for Ordinal Data of different dimensions

Participant: Christophe Biernacki.

This work has been motivated by a psychological survey on women affected by a breast tumor. Patients replied at different moments of their treatment to questionnaires with answers on ordinal scale. The questions relate to aspects of their life called dimensions. To assist the psychologists in analyzing the results, it is useful to emphasize a structure in the dataset. The clustering method achieves that by creating groups of individuals that are depicted by a representative of the group. From a psychological position, it is also useful to observe how questions may be grouped. This is why a clustering should be performed also on the features, which is called a co-clustering problem. However, gathering questions that are not related to the same dimension does not make sense from a psychologist stance. Therefore, the present work corresponds to perform a constrained co-clustering method aiming to prevent questions from different dimensions from getting assembled in a same column-cluster. In addition, evolution of co-clusters along time has been investigated. The method relies on a constrained Latent Block Model embedding a probability distribution for ordinal data. Parameter estimation relies on a Stochastic EM-algorithm associated to a Gibbs sampler, and the ICL-BIC criterion is used for selecting the numbers of co-clusters. The resulting work has been submitted to an international journal [42].

This is joint work with Margot Selosse and Julien Jacques, both from University of Lyon 2. Margot Selosse is a new PhD student co-supervised by Julien Jacques and Christophe Biernacki.

7.14. MASSICCC Platform for SaaS Software Availability

Participants: Christophe Biernacki, Vincent Kubicki, Jonas Renault, Josselin Demont, Matthieu Marbac.

MASSICCC is a demonstration platform giving access through a SaaS (service as a software) concept to data analysis libraries developed at Inria. It allows to obtain results either directly through a website specific display (specific and interactive visual outputs) or through an R data object download. It started in October 2015 for two years and is common to the Modal team (Inria Lille) and the Select team (Inria Saclay). In 2016, two packages have been integrated: Mixmod and MixtComp (see the specific section about MixtComp). In 2017, the BlockCluster package has been integrated and also a particular attention to provide meaningful graphical outputs (for Mixmod, MixtComp and BlockCluster) directly in the web platform itself has led to some specific developments.

MASSICCC has led to a second short meeting in February 2017 in Lille (after a first short meeting in April 2016 in Lille) for obtaining a feedback from company and academic users.

The MASSICCC platform is available here in the web: https://massiccc.lille.inria.fr

7.15. Model-Based Co-Clustering of Multivariate Functional Data

Participant: Christophe Biernacki.

High dimensional data clustering is an increasingly interesting topic in the statistical analysis of heterogeneous large-scale data. We consider the problem of clustering heterogeneous high-dimensional data where the individuals are described by functional variables which exhibit a dynamical longitudinal structure. We address the issue in the framework of model-based co-clustering and propose the functional latent block model (FLBM). The introduced FLBM model allows to simultaneously cluster a sample of multivariate functions into a finite set of blocks, each block being an association of a cluster over individuals and a cluster over functional variables. Furthermore, the homogeneous set within each block is modeled with a dedicated latent process functional regression model which allows its segmentation according to an underlying dynamical structure. The proposed model allows thus to fully exploit the structure of the data, compared to classical latent block clustering models for continuous non functional data, which ignores the functional structure of the observations. The FLBM can therefore serve for simultaneous co-clustering and segmentation of multivariate non-stationary functions. We propose a variational expectation-maximization (EM) algorithm (VEM-FLBM) to monotonically maximize a variational approximation of the observed-data log-likelihood for the unsupervised inference of the FLBM model. This work has been presented as an invited speaker to the 61th World Staistics Congress [30].

This is a joint work with Faicel Chamroukhi of University of Caen.

7.16. Reject Inference Methods in Credit Scoring: A Rational Review

Participants: Christophe Biernacki, Adrien Ehrhardt, Vincent Vandewalle.

The granting process of all credit institutions rejects applicants having a low credit score. Developing a scorecard, *i.e.* a correspondence table between a client's characteristics and his score, requires a learning dataset in which the target variable good/bad borrower is known. Rejected applicants are *de facto* excluded from the process. This biased learning population might have deep consequences on the scorecard relevance. Some works, mostly empirical ones, try to exploit rejected applicants in the scorecard building process. This work proposes a rational criterion to evaluate the quality of a scoring model for the existing Reject Inference methods and dig out their implicit mathematical hypotheses. It is shown that, up to now, no such Reject Inference method can guarantee a better credit scorecard. These conclusions are illustrated on simulated and real data from the french branch of Crédit Agricole Consumer Finance (CACF). An early version of this work has been presented as a talk in the national conference [31] and a preprint is being to be finalized.

This is a joint work with Philippe Heinrich of University of Lille and Sébastien Beben of Crédit Agricole Consumer Finance.

7.17. Survival Analysis with Complex Covariates: A Model-based Clustering Preprocessing Step

Participants: Christophe Biernacki, Vincent Vandewalle.

Many covariates are now available through sensors in the industrial context, and are expected to be related to the survival analysis target. Such covariates are often complex, what has to be understood as a possible mix between continuous, categorical, even functional over time, variables with the possibility to contain missing or uncertain values. A natural question in survival analysis is to design in both flexible and easy way an hazard function related to these potentially complex covariates, while preserving the opportunity to benefit from classical hazard functions.

In the context of a bilateral contract with Alstom company on the survival analysis topic, we have been invited to give a tutorial in the IEEE PHM International Conference on Prognostics and Health Management in US [22] . In this tutorial, we have described how to decompose the unknown targeted hazard function into two complementary parts. The first one can be any classical user hazard function conditional on a latent categorical variable. The second one is the distribution of this latent variable conditionally to the complex covariates. The way to combine both parts is to sum their product over the latent variable (marginal distribution), leading to the final targeted hazard function. The key to perform this approach is to focus on the latent variable definition which can be obtained with a model-based clustering approach dedicated to complex covariates. Beyond a selected review of recent methodologies dedicated to clustering, we have described in depth some related software to perform previous clustering methods. Some case studies have been also provided in an industrial context. At the end of the talk the practitioner is thus able to perform such clustering method to use it finally with its own hazard function.

7.18. Dealing with Missing Data Through Mixture Models

Participants: Christophe Biernacki, Vincent Vandewalle.

Many data sets have missing values, however the majority of statistical methods need a complete dataset to work. Thus, practitioners often use imputation or multiple imputations to complete the data as a pre-processing step. Mixture models can be used to naturally deal with missing data in an integrated way depending on the purpose. Especially, they can be used to classify the data or derive estimates for the distances. This work as been presented in an international conference [21].

7.19. Review on Mixture Modeling and High-dimensional Clustering

Participant: Christophe Biernacki.

Following the Journées d'Études en Statistique on 2014 in Frejus, on the topic "model choice and model aggregation" where two lectures have been given respectively on mixture model and on high dimensional clustering, a book has been published in 2017 including two chapters related to these talks (respectively [33] and [34]).

The second chapter is a joint work with Cathy Maugis-Rabusseau of INSA Toulouse.

7.20. Dealing with Missing Not at Random Values in Model-based Clustering

Participant: Christophe Biernacki.

Missing values are current in modern data sets. In many situations, making the simplifying hypothesis that they are missing at random is not realistic. However, it is very challenging to propose sensible models which address the underlying missing process. We make such proposals specific to the clustering context, namely making the assumption that missing values are missing at random conditionally to clusters, thus leading to a quite natural not missing at random marginal model. A working paper is in progress.

It is a joint work with Julie Josse of Ecole Polytechnique and Gilles Celeux of Inria Saclay - Île de France.

7.21. Dealing with Several Cluster Variables

Participant: Vincent Vandewalle.

In model based clustering of quantitative data it is often supposed that only one clustering variable explains the heterogeneity of all the others variables. However, when variables come from different sources, it is often unrealistic to suppose that the heterogeneity of the data can only be explained by one variable. If such an assumption is made, this could lead to a high number of clusters which could be difficult to interpret. A model based multi-objective clustering is proposed, is assumes the existence of several latent clustering variables, each one explaining the heterogeneity of the data on some clustering projection. In order to estimate the parameters of the model an EM algorithm is proposed, it mainly relies on a reinterpretation of the standard factorial discriminant analysis in a probabilistic way. The obtained results are projections of the data on some principal clustering components allowing some synthetic interpretation of the principal clusters raised by the data. The behavior of the model is illustrated on simulated and real data. This work as been presented in an international conference [24].

8. Bilateral Contracts and Grants with Industry

8.1. Florimond Desprez

Participant: Guillemette Marot.

Florimond Desprez is a company which breeds plant varieties and produces seeds, spreading its innovations across the different sectors of agriculture. This 2 months contract aimed at selecting candidate markers explaining the relationship between genotypes, organoleptic and nutritional qualities of chicory. It is a joint work with Quentin Grimonprez (InriaTech engineer).

8.2. Arcelor-Mittal

Participants: Christophe Biernacki, Vincent Vandewalle.

Arcelor-Mittal is a leader company in steel industry. This 11 months contract (which began in 2016) aims at optimizing predictive maintenance from mixed data (continuous, categorical, functional) provided by multiple sensors disseminated in steel production lines. Several thousands of sensors are simultaneously involved in this study, most of them providing functional (chronological) values.

It is a joint work with Quentin Grimonprez and Vincent Kubicki (InriaTech engineers).

8.3. Alstom

Participants: Christophe Biernacki, Benjamin Guedj, Vincent Vandewalle.

Alstom is is a world leader company in integrated transport systems. This 10 months contract aims at optimizing predictive maintenance in rail switches from complex data, in particular chronological ones.

It is a joint work with Etienne Goffinet (InriaTech engineer).

8.4. Vallourec

Participant: Christophe Biernacki.

Vallourec is a world leader in premium tubular solutions for the energy markets and for other demanding industrial applications. This 9 months contract (which began in 2016) aims at predicting quality of tubular connections from mixed data (continuous, categorical, functional).

It is a joint work with Etienne Goffinet and Vincent Kubicki (InriaTech engineers).

8.5. Running Care

Participant: Christophe Biernacki.

Running Care is a young company providing sport and medical coaching, and personalized healthy advices, for injury prevention. It is based on a mobile and watch app that collects sports and medical data to make them smart. This 8 months contract aims at predicting injury risks for the runner based on past runs and planned ones. It uses also many other available information that the runner can provide through the app.

It is a joint work with Quentin Grimonprez (InriaTech engineer).

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. Main partners of bilille

Participant: Guillemette Marot.

Bilille, the bioinformatics platform of Lille, officially gathers from Nov. 2015 a few bioinformaticians, biostatisticians and bioanalysts from the following teams:

EA2694 (Univ. Lille, CHRU, Inria)

FRABIO, FR3688 (Univ. Lille,CNRS)

CBP / GFS (Univ. Lille, CHRU)

TAG (Univ. Lille, CNRS, INSERM, Institut Pasteur de Lille)

U1167 (Univ. Lille, CHRU, INSERM et Institut Pasteur de Lille)

U1011 (Univ. Lille, INSERM)

UMR8198 (Univ. Lille, CNRS)

LIGAN PM (Univ. Lille, CNRS)

BONSAI (Inria, Univ. Lille, CNRS).

These last teams are thus the main partners of Modal concerning biostatistics for bioinformatics. Guillemette Marot is the co-head of the platform and works in close collaboration with the following people for the leadership of the scientific strategy related to the platform:

H. Touzet, BONSAI, UMR 9189 (co-head of bilille)

P. Touzet, UMR 8198 (deputy head of bilille)

V. Chouraki, U1167

M. Figeac, CBP / GFS

D. Hot, TAG

V. Leclère, Insitut Charles Viollette

M. Lensink, UMR 8576.

9.1.2. Collaborations of the year linked to bilille, the bioinformatics and bioanalysis platform

Participants: Guillemette Marot, Vincent Vandewalle.

Guillemette Marot and Vincent Vandewalle have supervised the data analysis part or support in biostatistics tools testing for the following research projects involving engineers from bilille (only the names of the principal investigators of the project are given even if several partners are sometimes involved in the project):

UMR 8576, E. Goulas, FLAM project

JPARC, M.H. David, AGI-HOX project

JPARC, M.C. Chartier-Harlin, RNA-Seq meta-analysis

U1003, D. Gkika, TRP canals screening

UMR 1167, F. Pinet, INCA-Network project.

9.1.3. Coordinator of the regional (Haut-De-France) project

Participant: Sophie Dabo-Niang.

Sophie Dabo-Niang is the coordinator of the regional (Haut-De-France) project "Bridging cell biomechanical phenotype and their biological expressions for Cancer diagnosis (STATE-CELL)". It is a project in partnership with Modal-Inria, LIMMS UMI 2820, LEM UMR 9221, Yncrea Hauts de France, INSERM U908. This project is submitted to I-SITE ULNE, SUSTAIN proposals 2017.

9.2. National Initiatives

9.2.1. Programme of Investments for the Future (PIA)

Bilille is a member of two PIA "Infrastructures en biologie-santé":

France Génomique (https://www.france-genomique.org/spip/?lang=en)

IFB, French Institute of Bioinformatics (https://www.france-bioinformatique.fr/en)

As leader of the platform, Guillemette Marot is thus involved in these networks.

9.2.2. RHU PreciNASH

Participant: Guillemette Marot.

Acronym: PreciNASH

Project title: Non-alcoholic steato-hepatitis (NASH) from disease stratification to novel therapeutic approaches

Coordinator: F. Pattou

Duration: 5 years

Partners: FHU Integra and Sanofi

PreciNASH, project coordinated by Pr. F. Pattou (UMR 859, EGID), aims at better understanding non alcoholic stratohepatitis (NASH) and improving its diagnosis and care. In this RHU, Guillemette Marot supervises a 2 years post-doc, as her team EA 2694 is member of the FHU Integra. EA 2694 is involved in the WP1 for the development of a clinical-biological model for the prediction of NASH. Other partners of the FHU are UMR 859, UMR 1011 and UMR 8199, these last three teams being part of the labex EGID (European Genomic Institute for Diabetes). Sanofi is the main industrial partner of the RHU PreciNASH. The whole project will last 5 years (2016-2021).

9.2.3. ANR

9.2.3.1. ANR ClinMine

Participants: Cristian Preda, Vincent Vandewalle.

ClinMine Project-2014-2017

ANR project (ANR TECSAN - Technologie de la santé)

Main coordinator of the project: Clarisse Dhaenens, CRIStAL, USTL

7 partners - EA 1046 (Maladie d'Alzheimer et pathologies vasculaires, Faculté de Médecine, Lille), EA 2694 (Centre d'Etudes et de Recherche en Informatique Médicale - Faculté de Médecine, Lille), MODAL (Inria LNE), Alicante (Entreprise), CHRU de Montpelier, GHICL (Groupe Hospitalier de l'Institut Catholique de Lille), CRIStAL, USTL.

9.2.3.2. ANR TheraSCUD2022

Participant: Guillemette Marot.

Acronym: TheraSCUD2022

Project title: Targeting the IL-20/IL-22 balance to restore pulmonary, intestinal and metabolic homeostasis after cigarette smoking and unhealthy diet

Coordinator: P. Gosset

Duration: 3 years

Partners: CIIL Institut Pasteur de Lille and UMR 1019 INRA Clermont-Ferrand

TheraSCUD2022, project coordinated by P. Gosset (Institut Pasteur de Lille), studies inflammatory disorders associated with cigarette smoking and unhealthy diet (SCUD). Guillemette Marot is involved in this ANR project as head of bilille platform, and will supervise 1 year engineer on integration of omic data. The duration of this project is 3 years (2017-2020).

9.2.4. Working groups

Sophie Dabo-Niang belongs to the following working groups:

- STAFAV (STatistiques pour l'Afrique Francophone et Applications au Vivant)
- ERCIM Working Group on computational and Methodological Statistics, Nonparametric Statistics Team
- Ameriska

Benjamin Guedj belongs to the following working groups (GdR) of CNRS:

- ISIS (local referee for Inria Lille - Nord Europe)
- MaDICS
- MASCOT-NUM (local referee for Inria Lille - Nord Europe).

Guillemette Marot belongs to the StatOmique working group.

9.2.5. Other initiatives

Participants: Serge Iovleff, Cristian Preda, Vincent Vandewalle.

Serge Iovleff is the head of the project CloHe granted in 2016 by the Mastodons CNRS challenge "Big data and data quality". The project is axed on the design of classification and clustering algorithms for mixed data with missing values with applications to high spatial resolution multispectral satellite image time-series. Website. Cristian Preda and Vincent Vandewalle are also members of the CloHe project.

9.3. International Initiatives

9.3.1. Inria Associate Teams Not Involved in an Inria International Labs

9.3.1.1. Equipes associées nord-européennes

Participants: Christophe Biernacki, Benjamin Guedj.

Benjamin Guedj and Christophe Biernacki pursue a two years collaboration as "Equipes associées nord-européennes" with the Irish team "INSIGHT". The Centre for Data Analytics INSIGHT is about the size of Inria Lille - Nord Europe and is the main Irish research facility in Statistics and Machine Learning. It is focused on the next generation of machine learning (ML) and statistics (Stat) algorithms that can operate on large-scale dynamic data. Nial FRIEL (NF) is the leader of the ML/Stat axis of INSIGHT, Brendan MURPHY (BM) is a professor. The topic of this project is to manage statistical models inflation by the mean of model clustering.

Benjamin Guedj and Christophe Biernacki visited NF and BM in Dublin once in 2017 to progress in the current collaboration.

9.3.1.2. EMC and CIMPA

Participant: Sophie Dabo-Niang.

EMS (European Mathematical Society): Sophie Dabo-Niang is a nominated member of EMS-CDC (Committee of Developing counties). She will be vice-chair of this committee in 2018.

CIMPA (International Center of Pure and Applied Mathematics): Sophie Dabo-Niang is a nominated member of CIMPA.

9.3.1.3. SIMERGE

Participants: Sophie Dabo-Niang, Serge Iovleff.

Sophie Dabo-Niang and Serge Iovleff are members of SIMERGE, a LIRIMA project-team (January 2015-December 2017). It includes researchers from Mistis (Inria Grenoble - Rhône-Alpes, France) and Inria-MODAL (Lille Nord de France), LERSTAD (Laboratoire d'Etudes et de Recherches en Statistiques et Développement, Université Gaston Berger, Sénégal), IRD (Institut de Recherche pour le Développement, Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, Dakar, Sénégal) and LEM lab (Lille Economie et Management, University of Lille). This project is submitted for renewal with a new partner (Institut Paster of Dakar, Senegal).

9.3.2. Inria International Partners

9.3.2.1. Informal International Partners

Participant: Benjamin Guedj.

Benjamin Guedj collaborates with Wouter Koolen (CWI, Netherlands), Peter Grünwald (CWI & Leiden University, Netherlands).

Benjamin Guedj collaborates with Olivier Wintenberger (KU, Denmark).

9.4. International Research Visitors

9.4.1. Visits to International Teams

Participant: Pascal Germain.

Pascal Germain will visit to "Groupe de recherche en apprentissage automatique de l'Université Laval" (Québec, Canada) to work with Professor François Laviolette from 11/12/2017 to 20/12/2017.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

Sophie Dabo-Niang has organized a session "Spatial econometrics" of the conference" at the *1st International Conference on Econometrics and Statistics*", June 15-16, 2017, Hong-Kong, China.

Sophie Dabo-Niang co-organises a session "Regression models under non i.i.d. settings" of the 10th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2017)
(http://www.cmstatistics.org/CMStatistics2017/), December 16-18, 2017, London.

Vincent Vandewalle organizes a session "Model-based clustering" of the conference" at the 10th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2017)"
(http://www.cmstatistics.org/CMStatistics2017/), December 16-18, 2017, London.

10.1.1.1. General Chair, Scientific Chair

Benjamin Guedj, Pascal Germain (both at Modal) and Francis Bach (SIERRA, Inria Paris) co-organize a NIPS 2017 workshop, called "(Almost) 50 shades of Bayesian learning: PAC-Bayesian trends and insights". A large audience is expected, and the workshop has a series of prestigious international speakers. See the website.

Benjamin Guedj is the organizer of the Modal team scientific seminar.

10.1.1.2. Member of the Organizing Committees

Alain Celisse was the co-head of the organizing committee of the *Journée Statistique Mathématique*, January the 8th 2017 at IHP.

Guillemette Marot was the co-head of the organizing committee of JOBIM 2017 (https://project.inria.fr/jobim2017/fr/).

Sophie Dabo-Niang was:

Member of the scientific committee of "*LICMA'17*", May 16-19, 2017, Beyrouth, Lebanon (http://www.licma.net/)

Member of the international organizing committee of AMU Commission on Women in Mathematics in Africa-African Women Mathematicians Association 2017 AMUCWMA - AWMA Workshop, July 7-8, 2017, Rabat, Morocco (http://fsr.um5.ac.ma/PACOM2017/WPACOM2017.php)

Vincent Vandewalle is member of the scientific animation cell of the bilille platform which has organized two thematic days in 2017:

Omics-driven genome annotation, Lille, March 31st 2017 (https://wikis.univ-lille1.fr/bilille/omics_annotation_2017)

Working with networks and pathways in molecular biology, Lille, November 20th 2017 (https://wikis.univ-lille1.fr/bilille/networks_2017)

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

Sophie Dabo-Niang was member of the local committee program of ISI (61st World Statistics Congress), 16-21, July, 2017, Marrakech, Morocco (http://payment.isi2017.org/committees/local-programme-committee-lpc/).

Vincent Vandewalle was member of the committee program of JOBIM 2017 (https://project.inria.fr/jobim2017/fr/).

10.1.2.2. Reviewer

Benjamin Guedj is a reviewer for NIPS 2017, AISTATS 2018, ALT 2018 and ICLR 2018, ICML 2018.

Pascal Germain: 6th International Conference on Learning Representations (ICLR 2018).

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

Sophie Dabo-Niang is member of *Revista Colombiana de Estadística*, 2015- –.

Christophe Biernacki is an Associate Editor of the North-Western European Journal of Mathematics (NWEJM).

10.1.3.2. Reviewer - Reviewing Activities

Benjamin Guedj acted as a reviewer for the following journal papers: Math Reviews, and the Annals of the Institute of Statistical Mathematics, Electronic Journal of Statistics.

Sophie Dabo acted as a reviewer for the following journal papers: Statistical Inference for Stochastic Processes, Computational Statistics and Data Analysis, Statistics, ESAIM: Probability and Statistics, Journal of Multivariate Analysis, Journal of Nonparametric statistics, Statistics and Probability Letters, Electronic journal of statistics, Metrika, Annals of Applied Statistics, Statistical Methods and Applications, ...

Christophe Biernacki acted as a reviewer for the following journal papers: Statistics and Computing (STCO), Psychometrika (PMET), Test, Knowledge and Information Systems (KAIS), Journal of Machine Learning Research (JMLR), Journal of Statistical Software (JSS), Neurocomputing (NEUCOM), Computational Statistics and Data Analysis (CSDA).

Alain Celisse acted as a reviewer for the following journal papers: Annals of Statistics, Bernoulli, JMLR.

Serge Iovleff acted as a reviewer for the following journal papers: Statistics and Computing and Neural Processing Letters.

Vincent Vandewalle acted as a reviewer for the following journal papers: Statistics in Medicine, Expert Systems With Applications, Computational Statistics, Journal de la SFdS.

10.1.4. Invited Talks

Sophie Dabo:

LICMA'17, Lebanese International Conference on Mathematics and Applications, Beyrouth, 16-19 May 2017. Nonparametric regression models for spatial data.

JEF 2017 4th Days of Econometrics for Finance, Rabat, 15-16 July, 2017.
https://sites.google.com/site/jefconference/home. Functional autoregressive spatial models.

Pan African Congress of Mathematicians 2017, Rabat, 3-7 July, 2017.

http://fsr.um5.ac.ma/PACOM2017/. Spatial prediction over spatio-functional data.

SIS 2017 Statistics and Data Science: New Challenges, New generations, Florence, 28-30 June 2017, http://www.fupress.com/archivio/pdf/3407_11724.pdf. Quasi-Maximum Likelihood Estimators For Functional Spatial Autoregressive Models.

Christophe Biernacki gave several invited talks in 2017:

61st World Statistics Congress, Marrakech (Morocco), 16-21 July 2017, http://payment.isi2017.org/ [30]

Classification Society Conference, Santa Clara (USA), 21-24 June 2017, https://academicaffairs.ucsc.edu/classification-society-conference/ [27]

XXIVèmes Rencontres de la Société Francophone de Classification, Lyon (France), 28-30 June 2017, http://polytech-sfc2017.univ-lyon1.fr/ [28]

StatLearn, Lyon (France), 7th April 2017, http://www.univ-lyon2.fr/culture-savoirs/podcasts/statlearn-2017-727260.kjsp?RH=podcasts, [20]

Talk to the seminar of the INRA of Jouy-en-Josas, June 12th 2017

Alain Celisse:

New assessment of the cross-validation performance, Montpellier, 24th November, 2017.

Change-point detection with kernels, IHP, 3rd October, 2017.

Cristian Preda:

C. Preda, P. Bastien (2017), Functional models applied to data from image analysis, 61st World Statistics Congress - ISI2017, Marrakech, Morocco, 16-21 July, 2017

A. Amarioarei, C. Preda, Scan statistics for some dependent models and applications, 17th Conference of the ASMDA, London, 6-9 June, 2017.

C. Preda, Gilbert Saporta's contributions to functional data analysis, 17th Conference of the ASMDA, London, UK, 6-9 June, 2017.

A. Amarioarei, C. Preda, Approximation for the scan statistics distribution of a three dimensional Poisson process, IMS China 2017, Nanning, China, June 28 - July 3, 2017.

Vincent Vandewalle gave several invited talks in 2017:

V. Vandewalle, C. Preda, Clustering categorical functional data. Application to medical discharge letters, 20th conference of the society of probability and statistics of Roumania, Brasov (Roumania), April 28, 2017.

V. Vandewalle, C. Biernacki, Dealing with missing data through mixture models, 154th ICB Seminar on "Statistics and clinical practice" Warsaw May 11, 2017.

V. Vandewalle, C. Biernacki, Survival analysis with complex covariates: a model-based clustering preprocessing step, IEEE PHM Dallas June 19th, 2017.

V. Vandewalle, Simultaneous dimension reduction and multi-objective clustering, IFCS Meeting Tokyo August 8th, 2017.

Benjamin Guedj gave several invited talks in 2017:

University College London (3/2017)

Université Paris 5 (3/2017)

Hélioparc (Pau, 3/2017)

Inria SequeL seminar (3/2017)

KU Leuven (5/2017)

Institut de Recherche en Informatique de Toulouse (9/2017)

Inria TAU seminar (10/2017)

Machine Learning in the Real world workshop at Criteo (11/2017).

10.1.5. Leadership within the Scientific Community

Benjamin Guedj is an elected member of the board of the French Statistical Society (SFdS). He is also deputy general secretary since June 2017.

Benjamin Guedj is a member of the board of AMIES, the French Agency fostering collaborations between mathematicians and the private sector.

Guillemette Marot is responsible of bilille, the bioinformatics and bioanalysis platform of Lille. More information about the platform is available at https://wikis.univ-lille1.fr/bilille/.

Christophe Biernacki is the president (since 1012) of the data mining and learning group of the French statistical association (SFdS, http://www.sfds.asso.fr/).

10.1.6. Scientific Expertise

Sophie Dabo-Niang is expert for the l'Oréal "Women in Science" award since 2014.

Christophe Biernacki acted as a President of a HCERES committee for research evaluation. He was also an elected member to the "Conseil National des Universités" (CNU) from October 2015 to September 2017. From October 2017, he is member of the "Commission d'Evaluation" (CE) of Inria.

Alain Celisse is reviewer for the annual Research Fellowship Competition of Cambridge.

Guillemette Marot reviewed one project as expert for the ANR.

10.1.7. Research Administration

Benjamin Guedj is a member of the scientific Council of the Laboratoire Paul Painlevé (Maths Department of the University of Lille).

Benjamin Guedj is an elected member of Inria's Evaluation Committee (CE).

Sophie Dabo is the head of the MeQAME research team of Laboratory LEM-CNRS 9221.

Christophe Biernacki is "Délégué Scientifique" of the Inria Lille center from June 2017.

Guillemette Marot was member of the Research Commission of the University of Lille 2 until December 2017.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Benjamin Guedj is teaching

Master: Machine learning, Theory and Algorithms, 10h, Université du Maine, Le Mans, France

Master: Machine learning, Theory and Algorithms, 30h, ISUP, Paris, France

Master: Machine learning, Theory and Algorithms, 20h, Université Pierre et Marie Curie, Paris, France

Guillemette Marot is teaching:

Licence: Biostatistics, 9h, L1, U. Lille Droit et Santé, France

Licence: Health care Informatics, 6h, L1, U. Lille Droit et Santé, France

Master: Biostatistics, 45h, M1, U. Lille Droit et Santé, France

Master: Supervised classification, 22h, M1, Polytech Lille, France

Master: Statistics for Human Genetics, 1h, U. Lille Droit et Santé, France

Doctorat: Data analysis with R, 14h, U. Lille Droit et Santé, France

Doctorat: Data analysis with R, 7h, COMUE Lille Nord de France, France

Doctorat: RNA-Seq analysis, 6h, U. Lille Droit et Santé, France

Sophie Dabo is teaching:

Licence: Probability, 24h, U. Lille 3, France

Master: Advanced Statistics, 24h, U. Lille 3, France

Master: Biostatistics Statistics, 40h, U. Lille 1, France

Master: Non-parametric Statistics, 24h, UGB, Senegal

Master: Sophie Dabo: Spatial Statistics, 24h, U. Lille 3, France

Christophe Biernacki is head of the M2 “Ingénierie Statistique et Numérique” (http://mathematiques.univ-lille1.fr/Formation/) at University Lille 1. He has also the following teaching activities at the same University:

Master: Coaching project, 10h, M1

Master: Data analysis, 97.5h, M2

Master: Coaching internship, 20h, M2

From September 2017, he is on secondment at Inria, without any teaching duty.

Cristian Preda is teaching:

L1: Probability, 40h, Polytech Lille

L1: Inferential Statistics, 50h, Polytech Lille

M1: Data Analysis, 40h, Polytech Lille

M2: Biostatistics, 12h, Polytech Lille

M2: Functional data analysis, 12h, U. Lille 1

Serge Iovleff is teaching:

DUT-S1: Mathématiques discrètes, TD, 68h

DUT-S1: Algèbre linéaire, TD, 32h

DUT-S2: Analyse et méthodes numériques, TD/TP, 56h

DUT-S3: Modélisation Mathématiques, TD, 24h

DUT-S4: R.O. et aide à la décision, TD, 32h

Master Mathématiques Appliquées, Statistique - Ingénierie Mathématique: Object Oriented programming, CTD, M1, U. Lille 1, 20h

Master Mathématiques Fondamentales: Statistics, CTD, M2, U. Lille 1, 24h

Vincent Vandewalle was in CRCT last year without any teaching.

10.2.2. Supervision

PhD: Jérémie Kellner, Gaussian processes and reproducing kernels, Université Lille 1, 11/12/2016, supervision: C. Biernacki, A. Celisse.

PhD: Emad Drwesh, Spatial Statistics in Discrete-Choice Models, University Lille 3, december, 11th, 2017, supervision: Sophie Dabo-Niang, Jérôme Foncel.

PhD: Mohamed Salem Ahmed, Contribution to spatial statistics and functional data analysis, University Lille 3, december, 12th, 2017, Sophie Dabo-Niang, Mohamed Attouch.

PhD in progress: Zied Gharbi (Contribution to Spatial autoregressive models), 2014, supervision: Sophie Dabo-Niang, Laurence Broze.

PhD in progress: Dang Khoi Pham (Planning and re-planning of nurses in an oncology department using a multi-objective and interdisciplinary approach), 2016, supervision: Alejandra Duenas, Christine Di Martinelli, Sophie Dabo-Niang. item PhD in progress: H. Sarter, Outils statistiques pour la sélection de variables et l’intégration de données "cliniques" et "omiques" : développement et application au registre EPIMAD, December 1st, 2016, supervision: C. Gower, G. Marot.

PhD in progress: Le Li, "PAC-Bayesian Online Clustering: theory and algorithms", iAdvize & Université d'Angers, since 11/2014, Benjamin Guedj, Sébastien Loustau.

PhD in progress: Arthur Leroy, "Machine learning algorithms to improve athletes' performance", INSEP, since 10/2017, supervision: Benjamin Guedj, Servane Gey, Jean-François Toussaint.

PhD in progress: Maxime Brunin, Etude du compromis entre précision statistique et temps de calcul, 1/10/2014, supervision: C. Biernacki, A. Celisse.

PhD in progress: Yaroslav Averyanov, New early stopping times and reproducing kernels, 1/10/2017, supervision: C. Preda, A. Celisse.

PhD in progress: Anne-Lise Bedenel, June 2015, supervision: Christophe Biernacki, Laetitia Jourdan.

PhD in progress: Adrien Ehrhardt, June 2016, supervision: Christophe Biernacki, Philippe Heinrich and Vincent Vandewalle.

PhD in progress: Margot Selosse, October 2017, Christophe Biernacki and Julien Jacques.

Serge Iovleff supervises dYawo Mamoua Kobara's Master thesis "*Estimation of a hierarchical Bayesian model for penalized regression*" at African Institute for Mathematical Sciences (AIMS), Senegal, 2017.

10.2.3. Juries

Guillemette Marot was examiner at the PhD defense of M. Canouil, Univ. Lille, September 29, 2017.

Sophie Dabo-Niang was member of the jury of the thesis of Hiba Alawieh, University of Lille 1, March 13th, 2017.

Sophie Dabo-Niang was referee and member of the thesis jury of Rim Ben Elouefi, INSA Rennes, September 5th, 2017.

Sophie Dabo-Niang was president of the thesis jury of Alban Mbinan Mbina, University of FranceVille (Gabon) and University of Lille 1, October 28th, Gabon (FranceVille).

Christophe Biernacki participated as a reviewer to 4 PhD theses and as an examinator to 1 PhD thesis. He was president of a recruitment committee for an assistant professor position.

10.3. Popularization

Pascal Germain gave a short talk vulgarizing a research topic for first year university students. Part of the presentation of the Painlevé Mathematic Laboratory (University of Lille), December, 1st 2017.

Christophe Biernacki has given about 10 talks during 2017 for institutions (Inria, universities...), companies and other related events. He organized also a first short meeting in February 2017 in Lille for obtaining a feedback from company and academic users about the MASSICCC platform developed by the Modal and Select teams (https://massiccc.lille.inria.fr/#/).

Alain Celisse gave a talk at the Meetup in Pau in January, 23rd 2017: Change-point detection with structured objects.

Benjamin Guedj gave a Meetup talk in Pau (3/2017) on quasi-Bayesian learning and an application to digits recognition.

11. Bibliography

Major publications by the team in recent years

[1] S. Arlot, A. Celisse. *Segmentation of the mean of heteroscedastic data via cross-validation*, in "Statistics and Computing", 2010, vol. 21, pp. 613–632

[2] P. BATHIA, S. IOVLEFF, G. GOVAERT. *An R Package and C++ library for Latent block models: Theory, usage and applications*, in "Journal of Statistical Software", 2016, https://hal.archives-ouvertes.fr/hal-01285610

[3] C. BIERNACKI, G. CELEUX, G. GOVAERT. *Exact and Monte Carlo Calculations of Integrated Likelihoods for the Latent Class Model*, in "Journal of Statistical Planning and Inference", 2010, vol. 140, pp. 2991-3002, https://hal.archives-ouvertes.fr/hal-00554344

[4] C. BIERNACKI, J. JACQUES. *A generative model for rank data based on an insertion sorting algorithm*, in "Computational Statistics and Data Analysis", 2013, vol. 58, pp. 162-176 [*DOI :* 10.1016/J.CSDA.2012.08.008], https://hal.archives-ouvertes.fr/hal-00441209

[5] C. BIERNACKI, J. JACQUES. *Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm*, in "Statistics and Computing", 2016, vol. 26, n^{o} 5, pp. 929-943, https://hal.inria.fr/hal-01052447

[6] A. CELISSE, J.-J. DAUDIN, L. PIERRE. *Consistency of maximum likelihood and variational estimators in stochastic block model*, in "Electronic Journal of Statistics", 2012, pp. 1847–1899, http://projecteuclid.org/handle/euclid.ejs

[7] M. GIACOFCI, S. LAMBERT-LACROIX, G. MAROT, F. PICARD. *Wavelet-based clustering for mixed-effects functional models in high dimension*, in "Biometrics", March 2013, vol. 69, n^{o} 1, pp. 31-40 [*DOI :* 10.1111/J.1541-0420.2012.01828.X], http://hal.inria.fr/hal-00782458

[8] J. JACQUES, C. PREDA. *Funclust: a curves clustering method using functional random variables density approximation*, in "Neurocomputing", 2013, vol. 112, pp. 164-171, https://hal.archives-ouvertes.fr/hal-00628247

[9] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Model-based clustering of Gaussian copulas for mixed data*, in "Communications in Statistics - Theory and Methods", December 2016, https://hal.archives-ouvertes.fr/hal-00987760

[10] V. VANDEWALLE, C. BIERNACKI, G. CELEUX, G. GOVAERT. *A predictive deviance criterion for selecting a generative model in semi-supervised classification*, in "Computational Statistics and Data Analysis", 2013, vol. 64, pp. 220-236, https://hal.inria.fr/inria-00516991

Publications of the year

Articles in International Peer-Reviewed Journals

[11] M.-S. AHMED, M. K. ATTOUCH, S. DABO-NIANG. *Binary functional linear models under choice-based sampling*, in "Econometrics and Statistics ", July 2017 [*DOI :* 10.1016/J.ECOSTA.2017.07.001], https://hal.inria.fr/hal-01654079

[12] P. ALQUIER, B. GUEDJ. *An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization*, in "Mathematical Methods of Statistics", 2017, https://hal.inria.fr/hal-01251878

[13] P. ALQUIER, B. GUEDJ. *Simpler PAC-Bayesian Bounds for Hostile Data*, in "Machine Learning", 2017, forthcoming, https://hal.inria.fr/hal-01385064

[14] E. Chazard, G. Ficheur, J.-B. Beuscart, C. Preda. *How to Compare the Length of Stay of Two Samples of Inpatients? A Simulation Study to Compare Type I and Type II Errors of 12 Statistical Tests*, in "Value in Health", July 2017, vol. 20, n^{o} 7, pp. 992 - 998 [*DOI :* 10.1016/j.jval.2017.02.009], https://hal.archives-ouvertes.fr/hal-01655463

[15] S. Dabo-Niang, A. Amiri. *Density estimation over spatio-temporal data streams*, in "Econometrics and Statistics ", September 2017 [*DOI :* 10.1016/j.ecosta.2017.08.005], https://hal.inria.fr/hal-01654081

[16] J. Dubois, V. Dubois, H. Dehondt, P. Mazrooei, C. Mazuy, A. A. Sérandour, C. Gheeraert, P. Guillaume, E. Baugé, B. Derudas, N. Hennuyer, R. Paumelle, G. Marot, J. S. Carroll, M. Lupien, B. Staels, P. Lefebvre, J. Eeckhoute. *The logic of transcriptional regulator recruitment architecture at cis -regulatory modules controlling liver functions*, in "Genome Research", June 2017, vol. 27, n^{o} 6, pp. 985 - 996 [*DOI :* 10.1101/gr.217075.116], https://hal.archives-ouvertes.fr/hal-01647846

[17] E. Hebbinckuys, J.-P. Marissal, C. Preda, V. Leclercq. *Assessing the burden of Clostridium difficile infections for hospitals*, in "Journal of Hospital Infection", September 2017, pp. 1-7 [*DOI :* 10.1016/j.jhin.2017.08.023], https://hal.archives-ouvertes.fr/hal-01655460

[18] M. Marbac, C. Biernacki, V. Vandewalle. *Model-based clustering of Gaussian copulas for mixed data*, in "Communications in Statistics - Theory and Methods", 2017, vol. 46, n^{o} 23, pp. 11635-11656, https://hal.archives-ouvertes.fr/hal-00987760

[19] C. Preda, A. Dermoune. *Parametrizations, fixed and random effects*, in "Journal of Multivariate Analysis", February 2017, vol. 154, pp. 162 - 176 [*DOI :* 10.1016/j.jmva.2016.11.001], https://hal.archives-ouvertes.fr/hal-01655461

Invited Conferences

[20] C. Biernacki, A. Lourme, M. Brunin, A. A. Celisse. *About Two Disinherited Sides of Statistics: Data Units and Computational Saving*, in "Statlearn 2017", Lyon, France, April 2017, pp. 1-56, https://hal.inria.fr/hal-01665905

[21] V. Vandewalle, C. Biernacki. *Dealing with missing data through mixture models*, in "ICB Seminars 2017 - 154th Seminar on ”Statistics and clinical practice”", Varsovie, Poland, May 2017, pp. 1-3, https://hal.inria.fr/hal-01667614

[22] V. Vandewalle, C. Biernacki. *Survival analysis with complex covariates: a model-based clustering preprocessing step*, in "IEEE PHM 2017", Dallas, United States, June 2017, https://hal.inria.fr/hal-01667588

[23] V. Vandewalle, T. Mottet, M. Marbac. *Model-based variable clustering*, in "CMStatistics/ERCIM 2017 - 10th International Conference of the ERCIM WG on Computational and Methodological Statistics", London, United Kingdom, December 2017, pp. 1-19, https://hal.inria.fr/hal-01691421

[24] V. Vandewalle. *Simultaneous dimension reduction and multi-objective clustering*, in "IFCS 2017 - Conference of the International Federation of Classification Societies", Tokyo, Japan, August 2017, pp. 1-29, https://hal.inria.fr/hal-01662271

Conferences without Proceedings

[25] M. BAELDE, C. BIERNACKI, R. GREFF. *A mixture model-based real-time audio sources classification method*, in "The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP2017", New Orleans, United States, March 2017, https://hal.archives-ouvertes.fr/hal-01420677

[26] M. BAELDE, C. BIERNACKI, R. GREFF. *Classification de signaux audio en temps-réel par un modèle de mélanges d'histogrammes*, in "JDS 2017 - 49e Journées de Statistiques", Avignon, France, May 2017, https://hal.archives-ouvertes.fr/hal-01592496

[27] C. BIERNACKI, A. LOURME. *Introduction Units in model-based clustering Units in model-based co-clustering Conclusion Unifying Data Units and Models in (Co-)Clustering*, in "2017 - Classification Society Conference", Santa Clara, United States, June 2017, pp. 1-38, https://hal.archives-ouvertes.fr/hal-01653896

[28] C. BIERNACKI, A. LOURME. *Units in model-based clustering Units in model-based co-clustering* , in "24e rencontres de la Société Francophone de Classification", Lyon, France, June 2017, https://hal.archives-ouvertes.fr/hal-01653899

[29] M. BRUNIN, C. BIERNACKI, A. CELISSE. *Compromis précision - temps de calcul appliqué au problème de régression linéaire*, in "2017 - 49e Journées de Statistique de la SFdS", Avignon, France, May 2017, pp. 1-6, https://hal.archives-ouvertes.fr/hal-01653754

[30] F. CHAMROUKHI, C. BIERNACKI. *Model-Based Co-Clustering of Multivariate Functional Data*, in "ISI 2017 - 61st World Statistics Congress", Marrakech, Morocco, July 2017, https://hal.archives-ouvertes.fr/hal-01653782

[31] A. EHRHARDT, C. BIERNACKI, V. VANDEWALLE, P. HEINRICH, S. BEBEN. *Réintégration des refusés en Credit Scoring*, in "49e Journées de Statistique", Avignon , France, May 2017, https://hal.archives-ouvertes.fr/hal-01653767

[32] S. IOVLEFF, M. FAUVEL, S. GIRARD, C. PREDA, V. VANDEWALLE. *Mixture Models with Missing data Classication of Satellite Image Time Series: QUALIMADOS: Atelier Qualité des masses de données scientiques*, in "Journées Science des Données MaDICS 2017", Marseille, France, June 2017, pp. 1-60, https://hal.archives-ouvertes.fr/hal-01649206

Scientific Books (or Scientific Book chapters)

[33] C. BIERNACKI. *Mixture models*, in "Choix de modèles et agrégation", J.-J. DROESBEKE, G. SAPORTA, C. THOMAS-AGNAN (editors), Technip, September 2017, https://hal.inria.fr/hal-01252671

[34] C. BIERNACKI, C. MAUGIS. *High-dimensional clustering*, in "Choix de modèles et agrégation, Sous la direction de J-J. DROESBEKE, G. SAPORTA, C. THOMAS-AGNAN Edition: Technip", September 2017, https://hal.archives-ouvertes.fr/hal-01252673

Other Publications

[35] M. BAELDE, C. BIERNACKI, R. GREFF. *Real-time Audio Classification based on Mixture Models*, March 2017, The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2017), Poster, https://hal.archives-ouvertes.fr/hal-01481934

[36] C. BIERNACKI, A. LOURME. *Unifying Data Units and Models in (Co-)Clustering*, December 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01653881

[37] S. BOUKA, S. DABO-NIANG, G. M. NKIET. *On estimation in a spatial functional linear regression model with derivatives*, March 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01516616

[38] B. GUEDJ, B. SRINIVASA DESIKAN. *Pycobra: A Python Toolbox for Ensemble Learning and Visualisation*, April 2017, working paper or preprint, https://hal.inria.fr/hal-01514059

[39] A. HIBA, N. WICKER, C. BIERNACKI. *Projection under pairwise distance controls*, December 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01420662

[40] J. JACQUES, C. BIERNACKI. *Model-Based Co-clustering for Ordinal Data*, January 2017, working paper or preprint, https://hal.inria.fr/hal-01448299

[41] M. MARBAC, V. VANDEWALLE. *A tractable Multi-Partitions Clustering*, January 2018, https://arxiv.org/abs/1801.07063 - working paper or preprint, https://hal.inria.fr/hal-01691417

[42] M. SELOSSE, J. JACQUES, C. BIERNACKI, F. COUSSON-GÉLIE. *Analyzing health quality survey using constrained co-clustering model for ordinal data and some dynamic implication*, November 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01643910

[43] M. SELOSSE, J. JACQUES, C. BIERNACKI. *ordinalClust: a package for analyzing ordinal data*, January 2018, working paper or preprint, https://hal.inria.fr/hal-01678800