IN PARTNERSHIP WITH:
CNRS

Université Lille 2

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2018

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:

B2.2.3. - Cancer
B9.5.6. - Data science
B9.6.3. - Economy, Finance
B9.6.5. - Sociology

1. Team, Visitors, External Collaborators

Research Scientists

Christophe Biernacki [Team leader, Inria, Senior Researcher, HDR]
Pascal Germain [Inria, Researcher]
Benjamin Guedj [Inria, Researcher]
Hemant Tyagi [Inria, Researcher, from Dec 2018]

Faculty Members

Alain Celisse [Univ de Lille, Associate Professor, HDR]
Serge Iovleff [Univ de Lille, Associate Professor]
Guillemette Marot [Univ de Lille, Associate Professor]
Cristian Preda [Univ de Lille, Professor, HDR]
Vincent Vandewalle [Univ de Lille, Associate Professor]
Sophie Dabo [Univ de Lille, Associate Professor]

Post-Doctoral Fellow

Fabien Laporte [Ecole polytechnique, from Jul 2018]

PhD Students

Yaroslav Averyanov [Inria]
Maxime Baelde [A-Volute]
Anne-Lise Bedenel [MeilleureAssurance]

Adrien Ehrhardt [CA CF]
Arthur Leroy [Univ René Descartes]
Le Li [iAdvize & Univ. dAngers, until November 2018]

Technical staff

Vincent Kubicki [Inria, from Oct 2018]
Bhargav Srinivasa Desikan [Inria, until Aug 2018]

Interns

Mohamed Boulkhir [Inria, from Jul 2018 until Aug 2018]
Wilfried Heyse [Inria, from Jun 2018 until Aug 2018]
Hugo Lair [Inria, from Mar 2018 until Jun 2018]
Florent Latimier [Inria, from Jun 2018 until Aug 2018]
Gael Letarte [Inria, from Sep 2018]
Asmita Poddar [Inria, from May 2018 until Jul 2018]
Ibrahima Souane [Inria, from Apr 2018 until Sep 2018]

Administrative Assistant

Anne Rejl [Inria]

Visiting Scientist

Michael Gallaugher [McMaster University, from Mar 2018 until May 2018]

External Collaborators

Julien Jacques [Univ Lumière, HDR]
Philippe Heinrich [Univ de Lille]

2. Overall Objectives

2.1. Context

In several respects, modern society has strengthened the need for statistical analysis, even if other related names are sometimes preferably used depending on methods, communities and applications, as data analysis, machine learning or artificial intelligence. The genesis comes from the easier availability of data thanks to technological breakthroughs (storage, transfer, computing), and are now so widespread that they are no longer limited to large human organizations. The more or less conscious goal of such data availability is the expectation to improving the quality of "since the dawn of time" statistical stories which are namely discovering new knowledge or doing better predictions. These both central tasks can be referred respectively as unsupervised learning or supervised learning, even if it is not limited to them or other names exist depending on communities. Somewhere, it pursues the following hope: "more data for better and more results".

However, today's data are more and more complex. They gather mixed type features (for instance continuous data mixed with categorical data), missing or partially missing (like intervals) items and numerous variables (high dimensional situation). As a consequence, the target "better and more results" of the previous adage (both words are important: "better" and also "more") could not be reached through a somewhat "handwork" way, but should inevitably rely on some theoretical formalization and guarantee. Indeed, data can be so numerous and so complex (data can live in quite abstract spaces) that the "empirical" statistician is quickly outdated. However, data being subject by nature to randomness, the probabilistic framework is a very sensible theoretical environment to serve as a general guide for modern statistical analysis.

2.2. Goals

Modal is a project-team working on today's complex data sets (mixed data, missing data, high-dimensional data), for classical statistical targets (unsupervised learning, supervised learning, regression, ...) with approaches relying on the probabilistic framework. This latter can be tackled through both model-based methods (as mixture models for a generic tool) and model-free methods (as probabilistic bounds on empirical quantities). Furthermore, Modal is connected to the real world by applications, typically with biological ones (some members have this skill) but many other ones are also considered since the application coverage of the Modal methodology is very large. It is also important to note that, in return, applications are often real opportunities for initiating academic questioning for the statistician (case of the Bilille platform and some bilateral contracts of the team).

From the academic communities point of view, Modal can be seen as belonging simultaneously to both the statistical learning and machine learning ones, as attested by its publications. Somewhere it is the opportunity to make a bridge between these two stochastic communities around a common but large probabilistic framework.

3. Research Program

3.1. Research axis 1: Unsupervised learning

Scientific locks related to unsupervised learning are numerous, concerning the clustering outcome validity, the ability to manage different kinds of data, the missing data questioning, the dimensionality of the data set, ... Many of them are addressed by the team, leading to publication achievements, often with a specific package delivery (sometimes upgraded as a software or even as a platform grouping several software). Because of the variety of the scope, it involves nearly all the permanent team members, often with PhD students and some engineers. The related works are always embedded inside a probabilistic framework, typically model-based approaches but also model-free ones like PAC-Bayes (PAC stands for Probably Approximately Correct), because such a mathematical environment offers both a well-posed problem and a rigorous answer.

3.2. Research axis 2: Performance assessment

One main concern of the Modal team is to provide theoretical justifications on the procedures which are designed. Such guarantees are important to avoid misleading conclusions resulting from any unsuitable use. The main ingredient in proving these guarantees is the use of the PAC framework, leading to finite-sample concentration inequalities. More precisely, contributions to PAC learning rely on the classical empirical process theory and the PAC-Bayesian theory. The Modal team exploits these non-asymptotic tools to analyze the performance of iterative algorithms (such as gradient descent), cross-validation estimators, online change-point detection procedures, ranking algorithms, matrix factorization techniques and clustering methods, for instance. The team also develops some expertise on the formal dynamic study of algorithms related to mixture models (important models used in the previous unsupervised setting), like degeneracy for EM algorithm or also label switching for Gibbs algorithm.

3.3. Research axis 3: Functional data

Mainly due to technological advances, functional data are more and more widespread in many application domains. Functional data analysis (FDA) is concerned with the modeling of data, such as curves, shapes, images or a more complex mathematical object, though as smooth realizations of a stochastic process (an infinite dimensional data object valued in a space of eventually infinite dimension; space of squared integrable functions, ...). Time series are an emblematic example even if it should not be limited to them (spectral data, spatial data, ...). Basically, FDA considers that data correspond to realizations of stochastic processes, usually assumed to be in a metric, semi-metric, Hilbert or Banach space. One may consider, functional independent or dependent (in time or space) data objects of different types (qualitative, quantitative, ordinal, multivariate,

time-dependent, spatial-dependent, ...). The last decade saw a dynamic literature on parametric or non-parametric FDA approaches for different types of data and applications to various domains, such as principal component analysis, clustering, regression and prediction.

3.4. Research axis 4: Applications motivating research

The fourth axis consists in translating real application issues into statistical problems raising new (academic) challenges for models developed in Modal team. Cifre Phds in industry and interdisciplinary projects with research teams in Health and Biology are at the core of this objective. The main originality of this objective lies in the use of statistics with complex data, including in particular ultra-high dimension problems. We focus on real applications which cannot be solved by classical data analysis.

4. Application Domains

4.1. Economic World

The Modal team applies it research to the economic world through CIFRE Phd supervision such as CACF (credit scoring), A-Volute (expert in 3D sound), Meilleur Taux (insurance comparator), ...It also has many contracts with companies such as Decathlon (world leader in sports equipment), Arcelor-Mittal (steel industry) or Alstom (integrated transport systems).

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.

5. Highlights of the Year

5.1. Highlights of the Year

- Hemant Tyagi has been recruited as CR in the team.
- Three new ANR funded (one began in 2018, two will start in 2019).
- One H2020 European project funded (started in November 2018).
- One-year EIT European project called SysBooster with ApSys and Nokia.
- Creation of a startup using MODAL's technology (MixtComp software).

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
- Publication: Pycobra: A Python Toolbox for Ensemble Learning and Visualisation
- 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. MixAll

Clustering using Mixture Models
KEYWORDS: Clustering - Clustering package - Generative Models
FUNCTIONAL DESCRIPTION: MixAll is a model-based clustering package for modelling mixed data sets. It has been engineered around the idea of easy and quick integration of any kind of mixture models for any kind of data, under the conditional independence assumption. Currently five models (Gaussian mixtures, categorical mixtures, Poisson mixtures, Gamma mixtures and kernel mixtures) are implemented. MixAll has the ability to natively manage completely missing values when assumed as random. MixAll is used as an R package, but its internals are coded in C++ as part of the STK++ library (www.stkpp.org) for faster computation.

- Participant: Serge Iovleff
- Partner: Université Lille 1
- Contact: Serge Iovleff
- URL: https://cran.r-project.org/web/packages/MixAll/

6.9. simerge

Statistical Inference for the Management of Extrem Risks, Genetics and Global epidemiology
KEYWORD: Biclustering
FUNCTIONAL DESCRIPTION: Allows to perform Co-Clustering on binary (Bernoulli) and counting variables (Poisson) using co-variables.

- Partner: Inria
- Contact: Serge Iovleff

6.10. Platforms

6.10.1. 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. Axis 1: 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 published in an international journal [12].

An extension of this work has been also presented to an international workshop.The idea is to use the data units principle as a way for (co-)clustering model enlargement.

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

7.2. Axis 1: 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 now published in the international journal [18]. This is joint work Julien Jacques from University of Lyon 2.

7.3. Axis 1: 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 also be performed 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 is now under revision in an international journal [54] and has been presented to an international conference [38]. The related R package ordinalClust has been also written and has led to a specific preprint [57].

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2, and Florence Cousson-Gélie from University Paul Valéry Montpellier 3.

7.4. Axis 1: Model-based co-clustering for mixed type data

Participant: Christophe Biernacki.

Over decades, a lot of studies have shown the importance of clustering to emphasize groups of observations. More recently, due to the emergence of high-dimensional datasets with a huge number of features, co-clustering techniques have emerged and proposed several methods for simultaneously producing groups of observations and features. By synthesizing the dataset in blocks (the crossing of a row-cluster and a column-cluster), this technique can sometimes summarize better the data and its inherent structure. The Latent Block Model (LBM) is a well-known method for performing a co-clustering. However, recently, contexts with features of different types (here called mixed type datasets) are becoming more common. Unfortunately, the LBM is not directly applicable on this kind of dataset. The present work extends the usual LBM to the so-called Multiple Latent Block Model (MLBM) which is able to handle mixed type datasets. The inference is done through a Stochastic EM-algorithm embedding a Gibbs sampler and model selection criterion is defined to choose the number of row and column clusters. This method was successfully used on simulated and real datasets. This work is available as a preprint [55] which has been submitted to an international journal. It has also led to the R package mixedClust which has been presented to an international workshop [56] and has led to a specific preprint [56].

An adaptation of this general principle to the specific case of mixing textual and continuous data has been also proposed and presented to a national conference [26], with an international audience.

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2.

7.5. Axis 1: Model-Based Co-clustering with Co-variables

Participant: Serge Iovleff.

This work has been motivated by an epidemiological and genetic survey of malaria disease in Senegal. Data were collected between 1990 and 2008. It is based on a latent block model taking into account the problem of grouping variables and clustering individuals by integrating information given by a set of co-variables. Numerical experiments on simulated data sets and an application on real genetic data highlight the interest of this approach. BEM algorithm is deduced and implemented in R package *simerge* and has led to a specific preprint [24].

7.6. Axis 1: Relaxing the Identically Distributed Assumption in Gaussian Co-Clustering for High Dimensional Data

Participant: Christophe Biernacki.

A co-clustering model for continuous data that relaxes the identically distributed assumption within blocks of traditional co-clustering is presented. The proposed model, although allowing more flexibility, still maintains the very high degree of parsimony achieved by traditional co-clustering. A stochastic EM algorithm along with a Gibbs sampler is used for parameter estimation and an ICL criterion is used for model selection. Simulated and real datasets are used for illustr1ation and comparison with traditional co-clustering. This work has led to a preprint

This is a joint work with Michael Gallaugher (PhD student) and Paul Mcnicholas, both from McMaster University (Canada). Michael Gallaughter visited the Modal for three months in 2018.

7.7. Axis 1: Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

A generic method is introduced to visualize in a Gaussian-like way, and onto R^2, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have a thorough overview of the potentially complex clustering result. An entropic measure allows us to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis. This work has been submitted to an international journal [12] and has also been presented to an international conference [41].

This is a joint work with Matthieu Marbac from ENSAI.

7.8. Axis 1: A targeted multi-partitions clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

Clustering is generally not a purpose by itself, because its results are mainly tools used by the statistician for another analysis. Indeed, in many applications, clusters are assessed from a set of observed variables, then these clusters are used to predict other variables which are used or not in clustering. Because the final objective of prediction is not considered during cluster analysis, there is no reason to obtain relevant clusters for the variables to predict. We present a unified approach which simultaneously performs cluster analysis and prediction. This method considers that the variables to clusters arise from a product of finite mixture models which provides multiple partition. Moreover, the variables to predict are considered to be independent of the variables to cluster given the partition. The predictions are achieved by a generalized linear model. Model selection is conducted by optimizing the BIC. This optimization is achieved with a modified version of the EM algorithm which performs model selection and maximum likelihood inference simultaneously. An early version of this work has been presented to an international conference [37].

It is a joint work with Matthieu Marbac from ENSAI and with Mohamed Sedki from Université Paris-Sud.

7.9. Axis 1: Co-clustering: A versatile way to perform clustering in high dimension

Participant: Christophe Biernacki.

Standard model-based clustering is known to be very efficient for low-dimensional data sets, but it fails for properly addressing high dimension (HD) ones, where it suffers from both statistical and computational drawbacks. In order to counterbalance this curse of dimensionality, some proposals have been made to take into account redundancy and features utility, but related models are not suitable for too many variables. We advocate that co-clustering, an unsupervised mixture model learning method to define simultaneously groups of rows (individuals) and groups of columns (variables) on a data matrix, is of particular interest to perform HD clustering of individuals even if it is not its primary mission. Indeed, column clustering is recast as a strategy to control the variance of the estimation, the model dimension being driven by the number of groups of variables instead of the number of variables itself. However, the statistical counterpart of this important variance reduction brings naturally some important model bias. The purpose is to access (first in an empirical manner) the trade-off bias-variance of the co-clustering strategy in scenarios involving HD fundaments (correlated variables, irrelevant variables). We show the ability of co-clustering to outperform simple mixture row-clustering, even if co-clustering clearly corresponds to a misspecified model situation, revealing a promising manner to efficiently address (very) HD clustering. An early version of this work has been presented to an international conference [36].

It is a joint work with Christine Keribin from Université Paris-Sud.

7.10. Axis 1: Dealing with missing data in model-based clustering through a MNAR model

Participants: Christophe Biernacki, Fabien Laporte.

Since the 90s, model-based clustering is largely used to classify data. Nowadays, with the increase of available data, missing values are more frequent. Traditional ways to deal with them consist in obtaining a filled data set, either by discarding missing values or by imputing them. In the first case, some information is lost; in the second case, the final clustering purpose is not taken into account through the imputation step. Thus, both solutions risk to blur the clustering estimation result. Alternatively, we defend the need to embed the missingness mechanism directly within the clustering modeling step. There exists three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In all situations logistic regression is proposed as a natural and flexible candidate model. In particular, its flexibility property allows us to design some meaningful parsimonious variants, as dependency on missing values or dependency on the cluster label. In this unified context, standard model selection criteria can be used to select between such different missing data mechanisms, simultaneously with the number of clusters. Practical interest of our proposal is illustrated on data derived from medical studies suffering from many missing data. An early version of this work has been presented to an international conference [33].

It is a joint work with Gilles Celeux from Inria Saclay and Julie Josse from Ecole Polytechnique.

7.11. Axis 1: Self Organizing Coclustering for textual data synthesis

Participant: Christophe Biernacki.

Recently, different studies have demonstrated the interest of co-clustering, which simultaneously produces clusters of lines and columns. The present work introduces a novel co-clustering model for parsimoniously summarizing textual data in documents × terms format. Besides highlighting homogeneous coclusters - as other existing algorithms do - we also distinguish noisy coclusters from significant ones, which is particularly useful for sparse documents × term matrices. Furthermore, our model proposes a structure among the significant coclusters and thus obtains a better interpretability to the user. By forcing a structure through row-clusters and column-clusters, this approach is competitive in terms of documents clustering, and offers user-friendly results. The algorithm derived for the proposed method is a Stochastic EM algorithm embedding a Gibbs sampling step and the Poisson distribution. A preprint is currently in progress.

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2.

7.12. Axis 1: Linking canonical and spectral clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

It is a recent work aiming at defining a mathematical bridge between classical model-based clustering and classical spectral clustering. Interest of such a prospect is to be able to compare both methods through the rigorous scheme of model selection paradigm. It is an ongoing work.

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

7.13. Axis 1: Multiple partition clustering

Participant: Vincent Vandewalle.

In the framework of model-based clustering, a model allowing several latent class variables have been proposed. This model assumes that the distribution of the observed data can be factorized into several independent blocks of variables. Each block is assumed to follow a latent class model (i.e., mixture with conditional independence assumption). The proposed model includes variable selection, as a special case, and is able to cope with the mixed-data setting. The simplicity of the model allows to estimate the repartition of the variables into blocks and the mixture parameters simultaneously, thus avoiding running EM algorithms for each possible repartition of variables into blocks. For the proposed method, a model is defined by the number of blocks, the number of clusters inside each block and the repartition of variables into blocks. Model selection can be done with two information criteria, the BIC and the MICL, for which an efficient optimization is proposed. The proposed method gives a rich interpretation of the data set at hand (i.e., analysis of the repartition of the variables into blocks and analysis of the clusters produced by each block of variables). This work as been presented in several international conferences and is now published [20].

It is a joint work with Matthieu Marbac from ENSAI.

7.14. Axis 2: Change-point detection by means of reproducing kernels

Participant: Alain Celisse.

Classical offline change-point detection approaches are limited to detecting changes arising in the mean and/or variance of the distribution along the time. Detecting changes in other moments of the distribution is possible, but at the price of stronger (unrealistic) distributional assumptions which are likely to be violated.

Reproducing kernels are a means to detect changes arising in any moments of the distribution along the time, which are not limited to the mean or the variance. One of the main contributions of this work is to provide a theoretically grounded model selection strategy allowing us to detection multiple changes. From additional extensive simulation experiments, it clearly arises that the so-called KCP approach outperforms numerous state-of-the art change-points detection procedures such as E-divisive, PELT, ...

7.15. Axis 2: New efficient algorithms for multiple change-point detection with kernels

Participants: Alain Celisse, Guillemette Marot.

Several statistical approaches based on reproducing kernels have been proposed to detect abrupt changes arising in the full distribution of the observations and not only in the mean or variance. Some of these approaches enjoy good statistical properties (oracle inequality, ...). Nonetheless, they have a high computational cost both in terms of time and memory. This makes their application difficult even for small and medium sample sizes ($n < 10^4$). This computational issue is addressed by first describing a new efficient and exact algorithm for kernel multiple change-point detection with an improved worst-case complexity that is quadratic in time and linear in space. It allows dealing with medium size signals (up to $n \approx 10^5$). Second, a faster but approximation algorithm is described. It is based on a low-rank approximation to the Gram matrix. It is linear in time and space. This approximation algorithm can be applied to large-scale signals ($n \geq 10^6$). These exact and approximation algorithms have been implemented in `R` and `C` for various kernels. The computational and statistical performances of these new algorithms have been assessed through empirical experiments. The runtime of the new algorithms is observed to be faster than that of other considered procedures. Finally, simulations confirmed the higher statistical accuracy of kernel-based approaches to detect changes that are not only in the mean. These simulations also illustrate the flexibility of kernel-based approaches to analyze complex biological profiles made of DNA copy number and allele B frequencies. An R package implementing the approach will be made available on github.

7.16. Axis 2: Multi-Layer Group-Lasso

Participants: Alain Celisse, Guillemette Marot.

Multi-Layer Group-Lasso (MLGL) is a new procedure of variable selection in the context of redundancy between explanatory variables, which holds true with high-dimensional data. A sparsity assumption is made that is, only a few variables are assumed to be relevant for predicting the response variable. In this context, the performance of classical Lasso-based approaches strongly deteriorate as the redundancy strengthens. The proposed approach combines variable aggregation and selection in order to improve interpretability and performance. First, a hierarchical clustering procedure provides at each level a partition of the variables into groups. Then, the set of groups of variables from the different levels of the hierarchy is given as input to group-Lasso, with weights adapted to the structure of the hierarchy. At this step, group-Lasso outputs sets of candidate groups of variables for each value of regularization parameter. The versatility offered by MLGL to choose groups at different levels of the hierarchy a priori induces a high computational complexity. MLGL however exploits the structure of the hierarchy and the weights used in group-lasso to greatly reduce the final time cost. The final choice of the regularization parameter – and therefore the final choice of groups – is made by a multiple hierarchical testing procedures. A paper associated to the R package MLGL has been submitted [45].

7.17. Axis 2: Pseudo-Bayesian Learning with Kernel Fourier Transform as Prior

Participants: Pascal Germain, Gael Letarte.

We revisit the kernel random Fourier features (RFF) method through the lens of the PAC-Bayesian theory. While the primary goal of RFF is to approximate a kernel, we look at the Fourier transform as a prior distribution over trigonometric hypotheses. It naturally suggests learning a posterior on these hypotheses. We derive generalization bounds that are optimized by learning a pseudo-posterior obtained from a closed-form expression, and corresponding learning algorithms This work has been accepted for publication at AISTATS 2019 conference [51].

It is a joint work with Emilie Morvant from Université Jean Monnet de Saint-Etienne.

7.18. Axis 2: Decentralized learning with budgeted network load using Gaussian copulas and classifier ensembles

Participant: Benjamin Guedj.

We examine a network of learners which address the same classification task but must learn from different data sets. The learners can share a limited portion of their data sets so as to preserve the network load. We introduce DELCO (standing for Decentralized Ensemble Learning with COpulas), a new approach in which the shared data and the trained models are sent to a central machine that allows to build an ensemble of classifiers. The proposed method aggregates the base classifiers using a probabilistic model relying on Gaussian copulas. Experiments on logistic regressor ensembles demonstrate competing accuracy and increased robustness as compared to gold standard approaches. A companion python implementation can be downloaded at https://github.com/john-klein/DELCO.

Joint work with John Klein, Olivier Colot, Mahmoud Albardan (all from CRIStAL lab, UMR 9189, Univ. Lille. Preprint submitted: [50].

7.19. Axis 2: Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly

Participants: Benjamin Guedj, Le Li.

When confronted with massive data streams, summarizing data with dimension reduction methods such as PCA raises theoretical and algorithmic pitfalls. Principal curves act as a nonlinear generalization of PCA and the present paper proposes a novel algorithm to automatically and sequentially learn principal curves from data streams. We show that our procedure is supported by regret bounds with optimal sublinear remainder terms. A greedy local search implementation that incorporates both sleeping experts and multi-armed bandit ingredients is presented, along with its regret bound and performance on a toy example and seismic data.

Preprint submitted: [47].

7.20. Axis 2: A Quasi-Bayesian Perspective to Online Clustering

Participants: Benjamin Guedj, Le Li.

When faced with high frequency streams of data, clustering raises theoretical and algorithmic pitfalls. We introduce a new and adaptive online clustering algorithm relying on a quasi-Bayesian approach, with a dynamic (i.e., time-dependent) estimation of the (unknown and changing) number of clusters. We prove that our approach is supported by minimax regret bounds. We also provide an RJMCMC-flavored implementation (called PACBO, see https://cran.r-project.org/web/packages/PACBO/index.html) for which we give a convergence guarantee. Finally, numerical experiments illustrate the potential of our procedure.

Joint work with Sébastien Loustau (LumenAI). Paper published in Electronic Journal of Statistics: https://projecteuclid.org/euclid.ejs/1537430425, [19].

7.21. Axis 2: Pycobra: A Python Toolbox for Ensemble Learning and Visualisation

Participants: Benjamin Guedj, Bhargav Srinivasa Desikan.

We introduce pycobra, a Python library devoted to ensemble learning (regression and classification) and visualisation. Its main assets are the implementation of several ensemble learning algorithms, a flexible and generic interface to compare and blend any existing machine learning algorithm available in Python libraries (as long as a predict method is given), and visualisation tools such as Voronoi tessellations. pycobra is fully scikit-learn compatible and is released under the MIT open-source license. pycobra can be downloaded from the Python Package Index (PyPi) and Machine Learning Open Source Software (MLOSS). The current version (along with Jupyter notebooks, extensive documentation, and continuous integration tests) is available at https://github.com/bhargavvader/pycobra and official documentation website is https://modal.lille.inria.fr/pycobra.

Paper published in Journal of Machine Learning Research: http://jmlr.org/papers/v18/17-228.html, [17]. Software submitted to the `scikit-learn-contrib` repository (under review).

7.22. Axis 2: Simpler PAC-Bayesian bounds for hostile data

Participant: Benjamin Guedj.

PAC-Bayesian learning bounds are of the utmost interest to the learning community. Their role is to connect the generalization ability of an aggregation distribution ρ to its empirical risk and to its Kullback-Leibler divergence with respect to some prior distribution π. Unfortunately, most of the available bounds typically rely on heavy assumptions such as boundedness and independence of the observations. This paper aims at relaxing these constraints and provides PAC-Bayesian learning bounds that hold for dependent, heavy-tailed observations (hereafter referred to as hostile data). In these bounds the Kullack-Leibler divergence is replaced with a general version of Csiszár's f-divergence. We prove a general PAC-Bayesian bound, and show how to use it in various hostile settings.

Joint work with Pierre Alquier (ENSAE ParisTech). Paper published in Machine Learning: [11].

7.23. Axis 2: PAC-Bayesian high dimensional bipartite ranking

Participant: Benjamin Guedj.

This paper is devoted to the bipartite ranking problem, a classical statistical learning task, in a high dimensional setting. We propose a scoring and ranking strategy based on the PAC-Bayesian approach. We consider nonlinear additive scoring functions, and we derive non-asymptotic risk bounds under a sparsity assumption. In particular, oracle inequalities in probability holding under a margin condition assess the performance of our procedure, and prove its minimax optimality. An MCMC-flavored algorithm is proposed to implement our method, along with its behavior on synthetic and real-life datasets.

Joint work with Sylvain Robbiano. Paper published in Journal of Statistical Planning and Inference: [16].

7.24. Axis 2: Multiview Boosting by Controlling the Diversity and the Accuracy of View-specific Voters

Participant: Pascal Germain.

We propose a boosting based multiview learning algorithm which iteratively learns i) weights over view-specific voters capturing view-specific information; and ii) weights over views by optimizing a PAC-Bayes multiview C-Bound that takes into account the accuracy of view-specific classifiers and the diversity between the views. We derive a generalization bound for this strategy following the PAC-Bayes theory which is a suitable tool to deal with models expressed as weighted combination over a set of voters. This work has been submitted to an international journal and is available as a preprint [44].

It is a joint work with Emilie Morvant from Université Jean Monnet de Saint-Etienne and with Massih-Reza Amini of Université de Grenoble, and with Anil Goyal affiliated to both institutions.

7.25. Axis 3: Clustering spatial functional data

Participants: Sophie Dabo, Cristian Preda, Vincent Vandewalle.

We propose two approaches for clustering spatial functional data. The first one is the model-based clustering that uses the concept of density for functional random variables. The second one is the hierarchical clustering based on univariate statistics for functional data such as the functional mode or the functional mean. These two approaches take into account the spatial features of the data: two observations that are spatially close share a common distribution of the associated random variables. The two methodologies are illustrated by an application to air quality data. This work will appear in the "Geostatistical Functional Data Analysis: Theory and Methods". Wiley, 2018. Editors : Jorge Mateu, Ramon Giraldo [39].

7.26. Axis 3: Categorical functional data analysis

Participants: Cristian Preda, Vincent Vandewalle.

We develop and implement techniques for analysis of categorical functional data. Visualization, clustering and regression methods with categorical functional predictor are proposed.

7.27. Axis 4: Real-time Audio Sources Classification

Participants: Christophe Biernacki, Maxime Baelde.

This work addresses the recurring challenge of real-time monophonic and polyphonic audio source classification. The whole power spectrum is directly involved in the proposed process, avoiding complex and hazardous traditional feature extraction. It is also a natural candidate for polyphonic events thanks to its additive property in such cases. The classification task is performed through a nonparametric kernel-based generative modeling of the power spectrum. Advantage of this model is twofold: it is almost hypothesis free and it allows to straightforwardly obtain the maximum a posteriori classification rule of online signals. Moreover it makes use of the monophonic dataset to build the polyphonic one. Then, to reach the real-time target, the complexity of the method can be tuned by using a standard hierarchical clustering preprocessing of sound models, revealing a particularly efficient computation time and classification accuracy trade-off. The proposed method reveals encouraging results both in monophonic and polyphonic classification tasks on benchmark and owned datasets, even in real-time situations. This method also has several advantages compared to the state-of-the-art methods include a reduced training time, no hyperparameters tuning, the ability to control the computation - accuracy trade-off and no training on already mixed sounds for polyphonic classification. This work is now under revision to an international journal [40].

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

7.28. Axis 4: Matching of descriptors evolving over time

Participants: Christophe Biernacki, Anne-Lise Bedenel.

In the web domain, and in particular for insurance comparison, data constantly evolve, implying that it is difficult to directly exploit them. For example, to do a classification, performing standard learning processes require data descriptors equal for both learning and test samples. Indeed, for answering web surfer expectation, online forms whence data come from are regularly modified. So, features and data descriptors are also regularly modified. In this work, it is introduced a process to estimate and understand connections between transformed data descriptors. This estimated matching between descriptors will be a preliminary step before applying later classical learning methods. This work has been presented to a national conference [27], with an international audience, and also to an international conference [28].

It is a joint work with Laetitia Jourdan, from University of Lille.

7.29. Axis 4: Supervised multivariate discretization and levels merging for logistic regression

Participants: Christophe Biernacki, Vincent Vandewalle, Adrien Ehrhardt.

For regulatory and interpretability reasons, the logistic regression is still widely used by financial institutions to learn the refunding probability of a loan given the applicants characteristics from historical data. Although logistic regression handles naturally both quantitative and qualitative data, three ad hoc pre-processing steps are usually performed: firstly, continuous features are discretized by assigning factor levels to predetermined intervals; secondly, qualitative features, if they take numerous values, are grouped; thirdly, interactions (products between two different features) are sparsely introduced. By reinterpreting these discretized (resp. grouped) features as latent variables and by modeling the conditional distribution of each of these latent variables given each original feature with a polytomous logistic link (resp. contingency table), a novel model-based resolution of the discretization problem is introduced. Estimation is performed via a Stochastic Expectation-Maximization (SEM) algorithm and a Gibbs sampler to find the best discretization (resp. grouping) scheme w.r.t. any classical logistic regression loss (AIC, BIC, test set AUC, ...). For detecting interacting features, the same scheme is used by replacing the Gibbs sampler by a Metropolis-Hastings algorithm. The good performances of this approach are illustrated on simulated and real data from Credit Agricole Consumer Finance. This work has been presenting to an international conference in statistics [35] and has been also submitting to an international conference in machine learning. [42].

This is a joint work with Philippe Heinrich from University of Lille.

7.30. Axis 4: MASSICCC Platform for SaaS Software Availability

Participant: Christophe Biernacki.

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 (forMixmod,MixtComp and BlockCluster) directly in the web platform itself has led to some specific developments. In 2018, MASSICCC has been presented to a workshop [29]. Currently, a preprint for an international journal dedicated to software is also in progress.

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

7.31. Axis 4: ClinMine: Optimizing the Management of Patients in Hospital

Participants: Cristian Preda, Vincent Vandewalle.

A better understanding of "patient pathway" thanks to data analysis can lead to better treatments for patients. The ClinMine project, supported by the French National Research Agency (ANR), aims at proposing, from various case studies, algorithmic and statistical models able to handle this type of pathway data, focusing primarily on hospital data.

Case studies, focusing on the integration of temporal data within analysis has been published [14]. First, the hypothesis that some aspects of the patient pathway can be described, even predicted, from the management process of the hospital medical mail is studied. Therefore a specific functional data analysis is driven, and several types of patients have been detected. The second case study deals with the detection of profiles through a biclustering of the patients. The difficulty to simultaneously deal with heterogeneous data, including temporal data is exposed and a method is proposed.

7.32. 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 R2 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 [48].

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

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts: SEMENCES DE FRANCE

Sophie Dabo-Niang has a contract with the enterprise SEMENCES DE FRANCE, concerning the realisation of a statistical software.

8.2. Bilateral Contracts: Arcelor-Mittal

Participants: Christophe Biernacki, Vincent Vandewalle.

Arcelor-Mittal is a leader company in steel industry. This contract (which began in 2016 and finished in 2018) 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. Bilateral Contracts: Alstom

Participants: Christophe Biernacki, Benjamin Guedj.

Alstom is a world leader company in integrated transport systems. This contract aims at optimizing predictive maintenance from free text annotations provided by maintenance people. The proposal consists in using co-clustering as a way for grouping both maintenance operations and words describing them.

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

8.4. Bilateral Contracts: Decathlon

Participant: Christophe Biernacki.

Decathlon is a leading sports retailer.

It is a joint work with Etienne Goffinet (InriaTech engineer). The purpose was to propose a innovative method for sales forecast by using complex data they have (mixed data, chronological series, etc.).

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. Bilille partnership

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)

C. Bellenguez, U1167

M. Figeac, CBP / GFS

D. Hot, TAG

V. Leclère, Insitut Charles Viollette

M. Lensink, UMR 8576.

O. Sand, IFB-Core.

9.1.2. Bilille collaborations

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):

CIIL, C. Faveeuw, Analysis of cytometry data

CIIL, P. Brodin, Analysis of phenotypic screening data

JPARC, J.M. Taymans, Analysis of translation chips

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

JPARC, A. Vincent, Microarray analysis

UMR 1167, F. Pinet, Analysis of proteomic data.

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

9.2.2. RHU PreciNASH

Participant: Guillemette Marot.

RHU PreciNASH

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

Abstract: 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 a 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. INS2I-CNRS project PEPS JCJC 2018 "PaRaFF"

Participant: Pascal Germain.

Projet PaRaFF: PAC-Bayesian Random Fourier Features

Coordinator: Emilie Morvant, Hubert Curien Lab, University Jean Monnet, Saint-Etienne

Year: 2018

Abstract: In data science, any method is based on a representation of the data. In this project, we study the learning of representation in the context of automatic learning methods called kernel methods. Our analysis is based on the Random Fourier Features, a method of approximating a kernel function based on a combination random attributes (combination defined by a probability distribution on the attributes). We aim to provide a theoretical understanding of this approach via PAC-Bayesian theory, and to propose a representation learning procedure by exploiting the specificities of this theory.

9.2.4. ANR

9.2.4.1. ANR APRIORI

Participants: Benjamin Guedj, Pascal Germain.

APRIORI 2019–2023, ANR PRC

PAC-Bayesian theory and algorithms for deep learning and representation learning.

Main coordinator of the project: Emilie Morvant, Université Jean Monnet.

Funding: 300k EUR.

2 partners - MODAL (Inria LNE), Hubert Curien Lab. (UMR CNRS 5516).

9.2.4.2. ANR BEAGLE

Participants: Benjamin Guedj, Pascal Germain.

BEAGLE 2019–2023, ANR JCJC

PAC-Bayesian theory and algorithms for agnostic learning

Main coordinator of the project: Benjamin Guedj

Funding: 180k EUR

The consortium also includes Pierre Alquier (ENSAE ParisTech), Peter Grünwald (CWI, The Netherlands), Rémi Bardenet (UMR CRIStAL 9189).

9.2.4.3. ANR SMILE

Participants: Christophe Biernacki, Vincent Vandewalle.

SMILE Project-2018-2022

ANR project (ANR SMILE - Statistical Modeling and Inference for unsupervised Learning at LargE-Scale)

Main coordinator of the project: Faicel Chamroukhi, LMNO, Université de Caen

4 partners - MODAL (Inria LNE), LMNO UMR CNRS 6139 (Caen), LMRS UMR CNRS 6085 (Rouen), LIS UMR CNRS 7020 (Toulon).

9.2.4.4. 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.

9.2.4.5. 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

Abstract: 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 bililleplatform, and will supervise 1 year engineer on integration of omic data. The duration of this project is 3 years (2017-2020).

9.2.5. 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.

9.2.6. 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. European Initiatives

9.3.1. FP7 & H2020 Projects

Benjamin Guedj and Vincent Vandewalle are involved on the European H2020 porject PERF-AI

Program: H2020

Project acronym: PERF-AI

Project title: Enhance Aircraft Performance and Optimisation through utilisation of Artificial Intelligence

Duration: November 2018 - November 2020.

Coordinator: Safety Line

Other partners: Safety Line

Abstract: PERF-AI will apply Machine Learning techniques on flight data (parametric and non-parametric approaches) to accurately measure actual aircraft performance throughout its lifecycle.

Within current airline operations, both at flight preparation (on-ground) and at flight management (in-air) levels, the trajectory is first planned, then managed by the Flight Management System (FMS) using a single manufacturer's performance model that is the same for every aircraft of the same type, and also on weather forecast that is computed long before the flight. It induces a lack of accuracy during the planning phase with a flight route pre-established at specific altitudes and speeds to optimize fuel burn, from take-off to landing using aircraft performances that are not those of the real aircraft. Also, the actual flight will usually shift from the original plan because of Air Traffic Control (ATC) constraints, adverse weather, wind changes and tactical re-routing, without possibility for the flight crew, either using the FMS or through connected services to tactically recompute the trajectory in order to continuously optimize the flight path. This is in particular due to the limitations of the performance databases that the current systems are using.

Hence, PERF-AI is focusing on identifying adequate machine learning algorithms, testing their accuracy and capability to perform flight data statistical analysis and developing mathematical models to optimize real flight trajectories with respect to the actual aircraft performance, thus, minimizing fuel consumption throughout the flight.

The consortium consists of Safety-Line (FR) and Inria (FR), having full expertise at Aircraft Performance and Data Science, hence, able to fully propose, test and validate different statistical models that will allow to accurately solve some optimization challenges and implement them in an operational environment.

9.3.2. Collaborations with Major European Organizations

Sophie Dabo-Niang is vice-chair of EMS-CDC (European Mathematical Society-Committe of Developing Countries). She is also a member of the executive committee of CIMPA (International Centre of Pure and Applied Mathematics)

Alain Celisse is a member of a one-year EIT European project called SysBooster with ApSys and Nokia.

9.4. International Initiatives

9.4.1. Inria International Labs

IIL CWI-Inria

Associate Team involved in the International Lab:

9.4.1.1. 6PAC

Title: Making Probably Approximately Correct Learning Active, Sequential, Structure-aware, Efficient, Ideal and Safe

International Partner (Institution - Laboratory - Researcher):

CWI (Netherlands) - Machine Learning Group

PI: Benjamin Guedj

Consortium: Peter Grünwald (co-PI), Wouter Koolen (CWI), Emilie Kaufmann (Inria LNE EPI SequeL).

Start year: 2018 (until 2021)

Webpage: https://bguedj.github.io/6pac/index.html

This project roots in statistical learning theory, which can be viewed as the theoretical foundations of machine learning. The most common framework is a setup in which one is given n training examples, and the goal is to build a predictor that would be efficient on new (similar) data. This efficiency should be supported by PAC (Probably Approximately Correct) guarantees, e.g. upper bounds on the excess risk of a predictor that hold with high probability. Such guarantees however often hold under stringent assumptions which are typically never met in real-life application, e.g., independent, identically distributed data. More realistic modelling of data has triggered many research efforts in several directions: first, accommodating possible data (e.g., dependent, heavy-tailed), and second, in the direction of sequential learning, in which the predictor can be built on the fly, while new data is gathered. We believe that an ever more realistic paradigm is active learning, a setup in which the learner actively requests data (possibly facing constraints, such as storage, velocity, cost, etc.) and adapts its queries to optimize its performance. The 3-years objective of 6PAC (where 6 stands for Sequential, Active, Efficient, Structured, Ideal, Safe - the six research directions we intend to contribute to) is to pave the way to new PAC generalization and sample-complexity upper and lower bounds beyond batch learning. Our ambition is to contribute to several learning setups, ranging from sequential learning (where data streams are collected) to adaptive and active learning (where data streams are requested by the learning algorithm).

9.4.2. Participation in Other International Programs

Starting December 2018, Benjamin Guedj is on sabbatical leave at University College London, Computer Science department, to lead a research team within the UCL AI center.

9.4.2.1. SIMERGE

Title: Statistical Inference for the Management of Extreme Risks and Global Epidemiology

International Partner (Institution - Laboratory - Researcher):

UGB (Senegal) - LERSTAD - Abdou Ka Diongue

Serge Iovleff and Sophie Dabo-Niang are associated members of SIMERGE.

9.5. International Research Visitors

9.5.1. Visits to International Teams

9.5.1.1. Research Stays Abroad

Sophie Dabo-Niang visited the University of Kuala Lumpur, November 2018, the University of Melbourne (Australia), December 2018, and the University of Nador (Morroco), end of December 2018.

Serge Iovleff visited several institutions in Senegal (April 22 – May 18, 2018). He gave a lecture at University Gaston Berger (UGB) of Saint-Louis and collaborated with Cheikh Loucoubar, Seydou Nourou Syllla and Cheikh Loucoubar of the team G4BBM of the Pasteur Institute of Dakar.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

Pascal Germain and Benjamin Guedj are the organizers of the https://modal.lille.inria.fr/wikimodal/doku.php?id=seminarsModal team scientific seminar.
Sophie Dabo-Niang is a co-organizer of the 2nd Conference on Econometrics for Environment.
Sophie Dabo-Niang, Cristian Preda and Vincent Vandewalle are the organizers of a session on "Functional Data Analysis" for the conference COMPSTAT 2018.
Vincent Vandewalle is the organizer of a session on advances in model based clustering for the conference ERCIM 2018.

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

Sophie Dabo-Niang is the chair of the Scientific Committee of CIMOM18.

10.1.2.2. Member of the Conference Program Committees

Christophe Biernacki is a member of the program committee of MBC2, an international workshop on Model-Based Clustering and Classification (http://mbc2.unict.it).

Cristian Preda was a member of the Scientific Committee of the 9th International Workshop on Applied Probability, IWAP 2018, 18-21 June Budapest, Hungary (https://iwap2018.com).

10.1.2.3. Reviewer

Pascal Germain acted as a reviewer for NIPS 2018, ICML 2018, ICLR 2018, CAp 2018.
Benjamin Guedj served as a reviewer for the top-tier conferences in machine learning ALT 2018, AISTATS 2018, NIPS 2018, ICML 2018, ICLR 2018. He also served as a reviewer for journals (Electronic Journal of Statistics, Journal of Machine Learning Research).
Sophie Dabo-Niang acted as a reviewer for JNP, JSPI, JRSSB, Spatial Statistics, Journal of SFDS, JMVA, ...
Christophe Biernacki acted as a reviewer for a dozen international statistical journals (CSDA, STCO, JMLR, IEEE PAMI, ...).
Serge Iovleff acted as a reviewer for Journal of Statistics and Computing.
Vincent Vandewalle acted as a reviewer for Statistics in Medicine, ADAC and Journal de la SFdS.
Alain Celisse acted as a reviewer for the Annals of Statistics, Bernoulli, JMLR, EJS, JSPI, Artificial Intelligence, ...
Cristian Preda acted as a reviewer for TEST, MCAP, JASA and Bernoulli.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

Christophe Biernacki is an Associate Editor of the North-Western European Journal of Mathematics (NWEJM) and for Frontiers on the topic "Computational Methods for Data Analytics". He is also a Guest Editor for the Special Issue on Innovations in Model-Based Clustering and Classification of the journal Advances Data Analysis and Classification (ADAC).

Cristian Preda is an Associate editor of the Methodology and Computing in Applied Probability Journal.

10.1.4. Invited Talks

Christophe Biernacki gave several invited talks:

- One talk at the international conference Compstat 2018
- Four talks at the Summer School on Clustering, Data Analysis and Visualization of Complex Data, May 2018, Catania, Italy [21] [22] [23] [30]
- Two talks at the Research Summer School on Statistics for Data Science – S4D, June 15th-22th 2018, Caen, France [32] [31]
- One talk at the international conference ERCIM 2018 [33]

Vincent Vandewalle:

- Invited talk at the international conference Compstat 2018 [25]
- Invited talk at the international conference ERCIM 2018
- Seminar of the EA 2694 (Université de Lille), Lille, France

Alain Celisse:

- ERCIM, Pise, 15 December 2018
- IWAP, Budapest, July 2018
- WeierstraßInstitute, Berlin, 2018

Pascal Germain:

- Journée Lilloise de Probabilité et Statistiques, Lille, France, 22 June, 2018
- Séminaire de l'équipe PS, Laboratoire Painlevé (University of Lille), Villeneuve d'Ascq France, 9 May, 2018

Benjamin Guedj:

- December 2018, GreekStochastics κ , Athens, Greece
- December 2018, 11th International Conference of the ERCIM working group on Computational and Methodological Statistics (CMStatistics 2018) (invited talk), Pisa, Italy
- September 2018, 2nd Italian-French Statistics Seminar (invited talk), Grenoble, France
- June 2018, 2nd annual congress of the French Mathematical Society (invited talk), Lille, France

Cristian Preda:

- One-Dimensional Discrete Scan Statistics Associated to Some Dependent Models, 5th Stochastic Modeling Techniques and Data Analysis International Conference. 12 - 16 June 2018, Chania, Crete, Greece

10.1.5. Leadership within the Scientific Community

Sophie Dabo-Niang is vice-chair of EMS-CDC.

Till May 2018, Christophe Biernacki was the president of the data mining and learning group of the French statistical association (SFdS, http://www.sfds.asso.fr/).

Since May 2018, Benjamin Guedj has served as president of the Machine Learning and Artificial Intelligence group (MALIA) of the French Statistical association (SFdS, http://www.sfds.asso.fr/).

Since 2017, Benjamin Guedj has been serving as a member of the boards of SFdS and AMIES.

10.1.6. Scientific Expertise

Guillemette Marot reviewed one project as an expert for the ANR and another one for ANSES.

10.1.7. Research Administration

Sophie Dabo-Niang is in charge of the MeQAME axis of the laboratory LEM, CNRS 9221.

Christophe Biernacki has been “Délégué Scientifique” of the Inria Lille center since June 2017.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Sophie Dabo-Niang is teaching

- Master: Spatial Statistics, 24h, M2, University of Lille, France
- Master: Advanced Statistics, 24h, M2, University of Lille, France
- Master: Multivaraite Data Analyses, 24h, M2, University of Lille, France
- Licence: Probability, 24h, L2, University of Lille, France
- Licence: Multivariate Statistics, 24h, L3, University of Lille, France

Guillemette Marot is teaching:

- Licence: Biostatistics, 12h, L1, University of Lille Droit et Santé, France
- Licence: Health care Informatics, 24h, L2, University of Lille Droit et Santé, France
- Master: Biostatistics, 45h, M1, University of Lille Droit et Santé, France
- Master: Supervised classification, 20h, M1, Polytech Lille, France
- Doctorat: Data analysis with R, 7h, University of Lille Droit et Santé, France
- Doctorat: RNA-Seq analysis, 12h, University of Lille Droit et Santé, France

Serge Iovleff is teaching

- Licence: Mathématiques discrètes, 68h, University of Lille, DUT Informatique
- Licence: Modélisation mathématique, 14h, University of Lille, DUT Informatique
- Licence: Algèbre linéaire, 32h, University of Lille, DUT Informatique
- Licence: Analyse et méthodes numériques, 56h, University of Lille, DUT Informatique
- Licence: R.O. et aide à la décision, 32h, University of Lille, DUT Informatique
- Formation Continue: Modélisation, 10h, University of Lille, DUT Informatique
- Master: Introduction to statistics, 16h, University of Lille

Cristian Preda is teaching

- Licence: Linear Regression, 24h, L3, University of Lille, France
- Master: Advanced Statistics, 24h, M1, University of Lille, France
- Master: Biostatistics, 10h, M2, University of Lille, France
- Master: Experimental Designs, 24h, M2, University of Lille, France

Alain Celisse is teaching

- Licence: Graphes et langages, 24h, L3, University of Lille, France
- Licence: Probabilités et statistique, 136h, L3, University of Lille, France
- Formation continue: Probabilités et statistique, 32h, L3, University of Lille, France

Pascal Germain is teaching

- Master: Introduction aux réseaux de neurones, 30h, M2, University of Lille, France

Benjamin Guedj is teaching

- Master: Bayesian Learning, 10h, M2, Centrale Lille, France

10.2.2. Supervision

PhD in progress: Yaroslav Averyanov, November 2017, supervision: Alain 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 defended: Le Li, November 2014–November 2018, supervision: Benjamin Guedj.

PhD in progress: Arthur Leroy, November 2017, supervision: Benjamin Guedj.

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

PhD in progress: Maxime Baelde, January 2016, Christophe Biernacki and Raphaël Greff.

PhD in progress: Hélène Sarter, Outils statistiques pour la sélection de variables et l'intégration de données “cliniques" et "omiques” : développement et application au registre EPIMAD, December 2016, Corinne Gower and Guillemette Marot.

10.2.3. Juries

Pascal Germain was an examiner at the PhD defense of Valentina Zantedeschi, University Jean Monnet of Saint-Etienne, December 18, 2018.
Benjamin Guedj served as a jury member for the PhD defense of Mahmoud Albardan, Univ. Lille, October 2018.
Sophie Dabo-Niang was a referee at the HDR defense of Tristan Kenga Kiesse, University of Rennes 1, October 22, 2018.
Sophie Dabo-Niang was a referee at the PhD defense of Ousmane Cisse, University Paris 1, December 11th, 2018, and of Julien Ndrin, University of Abidjan (Côte-d'Ivoire), April, 19th, 2018.
Sophie Dabo-Niang was a referee of the PhD dissertation of Javier Álvarez Lièbana, University of Granada (Spain), April, 2018.
Christophe Biernacki acted as a reviewer for PhD theses and one HdR defense. He also acted as an examinator for one PhD thesis and for one HdR defense.

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

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

Benjamin Guedj is an appointed deputy member of CLHSCT (Inria LNE).

Benjamin Guedj is an elected member of the Evaluation Committee (CE, Inria).

10.3.2. Internal action

- Pascal Germain gave a talk at the “30 minutes de sciences” seminar of Inria Lille (21/03/2018).
- Pascal Germain gave a talk vulgarizing his work at the “Café des sciences” of Inria Rocquencourt (13/11/2018).
- Guillemette Marot gave a talk at the “30 minutes de sciences” seminar of Inria Lille (21/12/2018).

11. Bibliography

Major publications by the team in recent years

[1] P. ALQUIER, B. GUEDJ. *Simpler PAC-Bayesian Bounds for Hostile Data*, in "Machine Learning", 2018 [*DOI :* 10.1007/s10994-017-5690-0], https://hal.inria.fr/hal-01385064

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

[4] C. BIERNACKI, A. LOURME. *Unifying Data Units and Models in (Co-)Clustering*, in "Advances in Data Analysis and Classification", May 2018, vol. 12, n° 41, https://hal.archives-ouvertes.fr/hal-01653881

[5] A. CELISSE. *Optimal cross-validation in density estimation with the L2-loss*, in "The Annals of Statistics", 2014, vol. 42, n° 5, pp. 1879–1910, https://hal.archives-ouvertes.fr/hal-00337058

[6] S. DABO-NIANG, C. TERNYNCK, A.-F. YAO. *Nonparametric prediction in the multivariate spatial context*, in "Journal of Nonparametric Statistics", 2016, vol. 28, n° 2, pp. 428-458 [*DOI :* 10.1080/10485252.2016.01.007], https://hal.inria.fr/hal-01425932

[7] 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° 6, pp. 985 - 996 [*DOI :* 10.1101/GR.217075.116], https://hal.archives-ouvertes.fr/hal-01647846

[8] 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

[9] M. MARBAC, V. VANDEWALLE. *A tractable Multi-Partitions Clustering*, in "Computational Statistics & Data Analysis", July 2018 [*DOI :* 10.1016/J.CSDA.2018.06.013], https://hal.inria.fr/hal-01691417

[10] 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

Publications of the year

Articles in International Peer-Reviewed Journals

[11] P. ALQUIER, B. GUEDJ. *Simpler PAC-Bayesian Bounds for Hostile Data*, in "Machine Learning", 2018 [*DOI :* 10.1007/S10994-017-5690-0], https://hal.inria.fr/hal-01385064

[12] C. BIERNACKI, A. LOURME. *Unifying Data Units and Models in (Co-)Clustering*, in "Advances in Data Analysis and Classification", May 2018, vol. 12, n° 41, https://hal.archives-ouvertes.fr/hal-01653881

[13] S. CURCEAC, C. TERNYNCK, T. B. OUARDA, F. CHEBANA, S. DABO-NIANG. *Short-term air temperature forecasting using Nonparametric Functional Data Analysis and SARMA models*, in "Environmental Modelling and Software", January 2019, vol. 111, pp. 394-408 [*DOI :* 10.1016/J.ENVSOFT.2018.09.017], https://hal.inria.fr/hal-01948928

[14] C. DHAENENS, J. JACQUES, V. VANDEWALLE, M. VANDROMME, E. CHAZARD, C. PREDA, A. AMARIOAREI, P. CHAIWUTTISAK, C. COZMA, G. FICHEUR, M.-E. KESSACI, R. PERICHON, J. TAILLARD, R. BORDET, A. LANSIAUX, L. JOURDAN, D. DELERUE, A. HANSSKE. *ClinMine: Optimizing the Management of Patients in Hospital*, in "IRBM", January 2018, vol. 39, n^{o} 2, pp. 83-92 [*DOI :* 10.1016/J.IRBM.2017.12.002], https://hal.inria.fr/hal-01692197

[15] R. GIRALDO, S. DABO-NIANG, S. MARTINEZ. *Statistical modeling of spatial big data: An approach from a functional data analysis perspective*, in "Statistics & Probability Letters", February 2018, vol. 136, pp. 126-129 [*DOI :* 10.1016/J.SPL.2018.02.025], https://hal.archives-ouvertes.fr/hal-01744181

[16] B. GUEDJ, S. ROBBIANO. *PAC-Bayesian High Dimensional Bipartite Ranking*, in "Journal of Statistical Planning and Inference", 2018 [*DOI :* 10.1016/J.JSPI.2017.10.010], https://hal.inria.fr/hal-01226472

[17] B. GUEDJ, B. SRINIVASA DESIKAN. *Pycobra: A Python Toolbox for Ensemble Learning and Visualisation*, in "Journal of Machine Learning Research", June 2018, vol. 18, pp. 1 - 5, https://hal.inria.fr/hal-01514059

[18] J. JACQUES, C. BIERNACKI. *Model-Based Co-clustering for Ordinal Data*, in "Computational Statistics & Data Analysis", July 2018, vol. 123, 15 p. , https://hal.inria.fr/hal-01448299

[19] L. LI, B. GUEDJ, S. LOUSTAU. *A Quasi-Bayesian Perspective to Online Clustering*, in "Electronic journal of statistics ", 2018 [*DOI :* 10.1214/18-EJS1479], https://hal.inria.fr/hal-01264233

[20] M. MARBAC, V. VANDEWALLE. *A tractable Multi-Partitions Clustering*, in "Computational Statistics & Data Analysis", July 2018 [*DOI :* 10.1016/J.CSDA.2018.06.013], https://hal.inria.fr/hal-01691417

Invited Conferences

[21] C. BIERNACKI. *Going further in cluster analysis and classification: Bi-clustering and co-clustering*, in "Summer School on Clustering, Data Analysis and Visualization of Complex Data", Catania, Italy, May 2018, https://hal.inria.fr/hal-01810380

[22] C. BIERNACKI. *Introduction to cluster analysis and classification: Evaluating clustering*, in "Summer School on Clustering, Data Analysis and Visualization of Complex Data", Catania, Italy, May 2018, https://hal.inria.fr/hal-01810377

[23] C. BIERNACKI. *Introduction to cluster analysis and classification: Performing clustering*, in "Summer School on Clustering, Data Analysis and Visualization of Complex Data", Catania, Italy, May 2018, https://hal.inria.fr/hal-01810376

[24] S. IOVLEFF, S. N. SYLLA. *blockcluster, simerge and C++ with R*, in "Mixture Models: Theory and Applications", Paris, France, June 2018, https://hal.inria.fr/hal-01884822

[25] V. VANDEWALLE, M. MARBAC. *A tractable multi-partitions clustering*, in "COMPSTAT 2018 - 23rd International Conference on Computational Statistics", Iasi, Romania, August 2018, https://hal.inria.fr/hal-01956922

National Conferences with Proceedings

[26] M. SELOSSE, J. JACQUES, C. BIERNACKI. *Co-clustering de données textuelles et continues*, in "SFdS 2018 - 50èmes Journées de Statistique", Saclay, France, May 2018, https://hal.inria.fr/hal-01797493

Conferences without Proceedings

[27] A.-L. BEDENEL, L. JOURDAN, C. BIERNACKI. *Probabilities estimation by a genetic algorithm*, in "ROADEF2018", Lorient, France, February 2018, https://hal.archives-ouvertes.fr/hal-01868195

[28] A.-L. BEDENEL, L. JOURDAN, C. BIERNACKI. *Probability estimation by an adapted genetic algorithm in web insurance*, in "LION 12 - Learning and Intelligent Optimization Conference", Kalamata, Greece, June 2018, https://hal.archives-ouvertes.fr/hal-01885117

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

[30] C. BIERNACKI. *Introduction to cluster analysis and classification: Formalizing clustering*, in "Summer School on Clustering, Data Analysis and Visualization of Complex Data", Catania, Italy, May 2018, https://hal.inria.fr/hal-01810379

[31] C. BIERNACKI. *Model selection theory and considerations in large scale scenarios*, in "Research Summer School on Statistics for Data Science – S4D", Caen, France, June 2018, https://hal.archives-ouvertes.fr/hal-01949168

[32] C. BIERNACKI. *Model-based clustering and co-clusteringin high-dimensional scenarios*, in "Research Summer School on Statistics for Data Science – S4D", Caen, France, June 2018, https://hal.archives-ouvertes.fr/hal-01949167

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

[34] C. BIERNACKI, V. VANDEWALLE, M. MARBAC. *Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering*, in "23rd International Conference on Computational Statistics", Iasi, Romania, August 2018, https://hal.archives-ouvertes.fr/hal-01949127

[35] A. EHRHARDT, V. VANDEWALLE, C. BIERNACKI, P. HEINRICH. *Supervised multivariate discretization and levels merging for logistic regression*, in "23rd International Conference on Computational Statistics", Iasi, Romania, August 2018, https://hal.archives-ouvertes.fr/hal-01949128

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

[37] M. MARBAC, C. BIERNACKI, M. SEDKI, V. VANDEWALLE. *A targeted multi-partitions clustering*, in "The 11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)", Pise, Italy, December 2018, https://hal.archives-ouvertes.fr/hal-01949111

[38] M. SELOSSE, J. JACQUES, C. BIERNACKI. *Analyzing large matrices of ordinal data*, in "The 11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)", Pise, Italy, December 2018, https://hal.archives-ouvertes.fr/hal-01949095

Scientific Books (or Scientific Book chapters)

[39] S. DABO-NIANG, C. PREDA, V. VANDEWALLE. *Clustering spatial functional data*, in "Geostatistical Functional Data Analysis : Theory and Methods. Editors: Jorge Mateu, Ramon Giraldo", 2018, https://hal.inria.fr/hal-01948934

Other Publications

[40] M. BAELDE, C. BIERNACKI, R. GREFF. *Real-Time Monophonic and Polyphonic Audio Classification from Power Spectra*, January 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01834221

[41] C. BIERNACKI, M. MARBAC, V. VANDEWALLE. *Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering*, December 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01949155

[42] A. EHRHARDT, C. BIERNACKI, V. VANDEWALLE, P. HEINRICH. *Feature quantization for parsimonious and interpretable predictive models*, December 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01949135

[43] P. GERMAIN, A. HABRARD, F. LAVIOLETTE, E. MORVANT. *PAC-Bayes and Domain Adaptation*, November 2018, https://arxiv.org/abs/1707.05712 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01563152

[44] A. GOYAL, E. MORVANT, P. GERMAIN, M.-R. AMINI. *Multiview Boosting by Controlling the Diversity and the Accuracy of View-specific Voters*, August 2018, https://arxiv.org/abs/1808.05784 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01857463

[45] Q. GRIMONPREZ, S. BLANCK, A. CELISSE, G. MAROT. *MLGL: An R package implementing correlated variable selection by hierarchical clustering and group-Lasso*, August 2018, working paper or preprint, https://hal.inria.fr/hal-01857242

[46] B. GUEDJ. *A Primer on PAC-Bayesian Learning*, January 2019, working paper or preprint, https://hal.inria.fr/hal-01983732

[47] B. GUEDJ, L. LI. *Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly*, May 2018, working paper or preprint, https://hal.inria.fr/hal-01796011

[48] A. HIBA, N. WICKER, C. BIERNACKI. *Projection under pairwise distance controls*, December 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01420662

[49] S. IOVLEFF, S. N. SYLLA, C. LOUCOUBAR. *Block clustering of Binary Data with Gaussian Co-variables*, December 2018, https://arxiv.org/abs/1812.08520 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01961978

[50] J. KLEIN, M. ALBARDAN, B. GUEDJ, O. COLOT. *Decentralized learning with budgeted network load using Gaussian copulas and classifier ensembles*, April 2018, https://arxiv.org/abs/1804.10028 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01779989

[51] G. LETARTE, E. MORVANT, P. GERMAIN. *Pseudo-Bayesian Learning with Kernel Fourier Transform as Prior*, 2018, https://arxiv.org/abs/1810.12683 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01908555

[52] G. MAZO, Y. AVERYANOV. *Constraining kernel estimators in semiparametric copula mixture models*, November 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01774629

[53] A. PODDAR, S. IOVLEFF, F. LATIMIER. *Estimation of Parsimonious Covariance Models for Gaussian Matrix Valued Random Variables for Multi-Dimensional Spectroscopic Data*, December 2018, WiML 2018 - 13th Women in Machine Learning workshop, Poster, https://hal.archives-ouvertes.fr/hal-01954769

[54] M. SELOSSE, J. JACQUES, C. BIERNACKI, F. COUSSON-GÉLIE. *Analyzing quality of life survey using constrained co-clustering model for ordinal data and some dynamic implication*, July 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01643910

[55] M. SELOSSE, J. JACQUES, C. BIERNACKI. *Model-based co-clustering for mixed type data*, October 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01893457

[56] M. SELOSSE, J. JACQUES, C. BIERNACKI. *mixedClust: an R package for mixed data classiication, clustering and co-clustering*, July 2018, 25th Summer Session Working Group on Model-Based Clustering, Poster, https://hal.archives-ouvertes.fr/hal-01949171

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