
IN PARTNERSHIP WITH:
**CNRS**

**Université Lille 2**

**Université des sciences et technologies de Lille (Lille 1)**

Activity Report 2016

# 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:**

3.1.4. - Uncertain data
3.2.3. - Inference
3.3.2. - Data mining
3.3.3. - Big data analysis
3.4.1. - Supervised learning
3.4.2. - Unsupervised learning
3.4.5. - Bayesian methods
3.4.7. - Kernel methods
5.2. - Data visualization
6.2.3. - Probabilistic methods
6.2.4. - Statistical methods
6.3.3. - Data processing
8.2. - Machine learning

**Other Research Topics and Application Domains:**

1.1.6. - Genomics
2.2.3. - Cancer
9.4.5. - Data science
9.5.3. - Economy, Finance
9.5.5. - Sociology

# 1. Members

**Research Scientist**

Benjamin Guedj [Inria, Research Scientist, Researcher]

**Faculty Members**

Christophe Biernacki [Team leader, Univ. Lille I, Faculty Member, Professor, HDR]
Alain Celisse [Univ. Lille I, Faculty Member, Associate Professor]
Serge Iovleff [Univ. Lille I, Faculty Member, Associate Professor]
Guillemette Marot [Univ. Lille II, Faculty Member, Associate Professor]
Cristian Preda [Univ. Lille I, Faculty Member, Professor, HDR]
Vincent Vandewalle [Univ. Lille II, Faculty Member, Associate Professor]

**Engineers**

Quentin Grimonprez [PhD Student funded by Univ. Lille 1 until Sep 2016, then Inria Engineer funded by Région Hauts-de-France]
Etienne Goffinet [Inria, granted by Région Hauts-de-France]
Vincent Kubicki [Inria, granted by Région Hauts-de-France]
Matthieu Marbac Lourdelle [Inria, from Dec 2016]

**PhD Students**

Maxime Baelde [A-Volute, granted by CIFRE]

Anne-Lise Bedenel [PIXEO, granted by CIFRE]
Maxime Brunin [Univ. Lille I]
Adrien Ehrhardt [CACF, from Apr 2016, granted by CIFRE]
Jeremie Kellner [Univ. Lille I, until Sep 2016]
Le Li [iAdvize, granted by CIFRE]

**Post-Doctoral Fellows**

Alexandru Amarioarei [Inria, granted by ANR CLINMINE project]
Cristina Preda [Inria, until Aug 2016]

**Visiting Scientist**

Emmanuel Chazard [Univ. Lille II, until Jul 2016]

**Administrative Assistant**

Anne Rejl [Inria]

**Others**

Samuel Blanck [Univ. Lille II, External Member]
Miguel Assuncao [Inria, Intern, from Sep 2016]
Rohit Uttam Bhagwat [Inria, Intern, from Jun 2016 until Aug 2016]
Faicel Chamroukhi [Univ. Caen, External Member, from Sep 2016, HDR]
Olivier Delrieu [C4X Discovery, External Member]
Philippe Heinrich [Univ. Lille I, External Member, from Jul 2016]
Hamza Cherkaoui [Inria, Intern, from May 2016 until Aug 2016]
Sophie Dabo [Univ. Lille III, External Member, HDR]
Astha Gupta [Inria, Intern, from May 2016 until Jul 2016]
Julien Jacques [Univ. Lyon II, External Member, HDR]
Siddharth Sharma [Inria, Intern, until May 2016]
Bhargav Srinivasa Desikan [Inria, Intern, from Aug 2016]

# 2. Overall Objectives

## 2.1. Overall Objectives

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

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

# 3. Research Program

## 3.1. Generative model design

The first objective of MODAL consists in designing, analyzing, estimating and evaluating new generative parametric models for multivariate and/or heterogeneous data. It corresponds typically to continuous and categorical data but it includes also other widespread ones like ordinal, functional, ranks,...Designed models have to take into account potential correlations between variables while being (1) justifiable and realistic, (2) meaningful and parsimoniously parameterized, (3) of low computational complexity. The main purpose is

to identify a few theoretical and general principles for model generation, loosely dependent on the variable nature. In this context, we propose two concurrent approaches which could be general enough for dealing with correlation between many types of homogeneous or heterogeneous variables:

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

## 3.2. Data visualization

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

# 4. Application Domains

## 4.1. Multiple domains applications

**Participants:** Sophie Dabo, Cristian Preda, Vincent Vandewalle, Alain Celisse, Benjamin Guedj, Christophe Biernacki, Guillemette Marot.

Modal targets a wide spectrum of application domains.

In particular, several members are interested in classification of functional data and functional regression models when data are correlated (temporally or spatially) and application to hydrological, environnemental or medical data.

Other topics include any application domains involving clustering, prediction or visualization (such as image segmentation, (online) clustering in retail, failure prediction in the steel industry, sales prediction in retail, ...). In most cases, we enforce the use of probabilistic models with associated software.

## 4.2. Genomics

**Participants:** Guillemette Marot, Alain Celisse.

With the use of high throughput technologies, more and more data are generated in molecular biology studies. Our developments are applied at several levels:

- genomics to detect aberrations in genomic profiles from patients suffering from cancers
- transcriptomics to find differentially expressed genes, e.g. between ill and healthy patients
- epigenetics to better understand cells mechanisms

# 5. Highlights of the Year

## 5.1. Highlights of the Year

The major highlight of Modal is related to transfer of its research towards the private sector. In 2016, several major bilateral contracts have been signed between Modal and leading international companies based in Hauts-de-France. Those collaborations directly proceed from the fundamental research carried within the team (see Section "Bilateral Contracts and Grants with Industry").

# 6. New Software and Platforms

## 6.1. BlockCluster

Block Clustering
KEYWORDS: Statistic analysis - Clustering package
SCIENTIFIC DESCRIPTION

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

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

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

## 6.2. Clustericat

FUNCTIONAL DESCRIPTION

Clustericat is an R package for model-based clustering of categorical data. In this package, the Conditional Correlated Model (CCM), published in 2014, takes into account the main conditional dependencies between variables through extreme dependence situations (independence and deterministic dependence). Clustericat performs the model selection and provides the best model according to the BIC criterion and the maximum likelihood estimates.

- Participants: Matthieu Marbac-Lourdelle, Vincent Vandewalle and Christophe Biernacki
- Contact: Matthieu Marbac-Lourdelle
- URL: https://r-forge.r-project.org/R/?group_id=1803

## 6.3. CoModes

FUNCTIONAL DESCRIPTION

CoModes is another R package for model-based clustering of categorical data. In this package, the Conditional Modes Model (CMM) (published in 2016) takes into account the main conditional dependencies between variables through particular modality crossings (so-called modes). CoModes performs the model selection and provides the best model according to the exact integrated likelihood criterion and the maximum likelihood estimates.

- Participants: Matthieu Marbac-Lourdelle, Vincent Vandewalle and Christophe Biernacki
- Contact: Matthieu Marbac-Lourdelle
- URL: https://r-forge.r-project.org/R/?group_id=1809

## 6.4. CorReg

FUNCTIONAL DESCRIPTION

The main idea of the CorReg package is to consider some form of sub-regression models, some variables defining others. We can then remove temporarily some of the variables to overcome ill-conditioned matrices inherent in linear regression and then reinject the deleted information, based on the structure that links the variables. The final model therefore takes into account all the variables but without suffering from the consequences of correlations between variables or high dimension.

- Participants: Clément Thery and Christophe Biernacki
- Contact: Clément Thery
- URL: https://cran.r-project.org/web/packages/CorReg/index.html

## 6.5. FunFEM

FUNCTIONAL DESCRIPTION

FunFEM package for R proposes a clustering tool for functional data. The model-based algorithm clusters the functional data into discriminative subspaces.

- Participants: Charles Bouveyron and Julien Jacques
- Contact: Charles Bouveyron
- URL: https://cran.r-project.org/web/packages/funFEM/index.html

## 6.6. FunHDDC

FUNCTIONAL DESCRIPTION

FunHDDC package for R proposes a clustering tool for functional data. The model-based clustering algorithm considers that functional data live in cluster-specific subspaces.

- Participants: Charles Bouveyron and Julien Jacques
- Contact: Charles Bouveyron
- URL: https://cran.r-project.org/web/packages/funHDDC/index.html

## 6.7. Galaxy - MPAgenomics

KEYWORDS: Bioinformatics - Data mining - Statistics - Genomics
FUNCTIONAL DESCRIPTION

Galaxy is an open, web-based platform for data intensive biomedical research. Galaxy features user friendly interface, workflow management, sharing functionalities and is widely used in the biologist community. The MPAgenomics R package developped by MODAL has been integrated into Galaxy, and the Galaxy MODAL instance has been publicly deployed thanks to the IFB-cloud infrastructure.

- Participants: Guillemette Marot and Samuel Blanck
- Contact: Guillemette Marot
- URL: https://cloud.france-bioinformatique.fr/accounts/login/

## 6.8. HDPenReg

High-Dimensional Penalized Regression
FUNCTIONAL DESCRIPTION

HDPenReg is an R-package based on a C++ code dedicated to the estimation of regression model with l1-penalization.

- Participants: Quentin Grimonprez and Serge Iovleff
- Contact: Quentin Grimonprez
- URL: https://cran.r-project.org/web/packages/HDPenReg/index.html

## 6.9. MPAGenomics

Multi-Patient Analysis of Genomic markers
KEYWORDS: Segmentation - Genomics - Marker selection - Biostatistics
SCIENTIFIC DESCRIPTION

MPAgenomics is an R package for multi-patients analysis of genomics markers. It enables to study several copy number and SNP data profiles at the same time. It offers wrappers from commonly used packages to offer a pipeline for beginners in R. It also proposes a special way of choosing some crucial parameters to change some default values which were not adapted in the original packages. For multi-patients analysis, it wraps some penalized regression methods implemented in HDPenReg.
FUNCTIONAL DESCRIPTION

MPAgenomics provides functions to preprocess and analyze genomic data. It is devoted to: (i) efficient segmentation and (ii) genomic marker selection from multi-patient copy number and SNP data profiles.

- Participants: Quentin Grimonprez, Guillemette Marot and Samuel Blanck
- Contact: Guillemette Marot
- URL: https://cran.r-project.org/web/packages/MPAgenomics/index.html

## 6.10. MetaMA

Meta-analysis for MicroArrays
KEYWORDS: Transcriptomics - Meta-analysis - Differential analysis - Microarrays - Biostatistics
FUNCTIONAL DESCRIPTION

MetaMA is a specialised software for microarrays. It is an R package which combines either p-values or modified effect sizes from different studies to find differentially expressed genes. The main competitor of metaMA is geneMeta. Compared to geneMeta, metaMA offers an improvement for small sample size datasets since the corresponding modelling is based on shrinkage approaches.

- Participant: Guillemette Marot
- Contact: Guillemette Marot
- URL: https://cran.r-project.org/web/packages/metaMA/index.html

## 6.11. MetaRNASeq

KEYWORDS: Transcriptomics - Meta-analysis - Differential analysis - High throughput sequencing - Biostatistics
FUNCTIONAL DESCRIPTION

This is joint work with Andrea Rau (INRA, Jouy-en-Josas). MetaRNASeq is a specialised software for RNA-seq experiments. It is an R package which is an adaptation of the MetaMA package presented previously. Both implement the same kind of methods but specificities of the two types of technologies require some adaptations to each one.

- Participants: Guillemette Marot and Andrea Rau
- Contact: Guillemette Marot
- URL: https://cran.r-project.org/web/packages/metaRNASeq/index.html

## 6.12. 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 (http://www.stkpp.org) for faster computation.

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

## 6.13. MixCluster

FUNCTIONAL DESCRIPTION

MixCluster is an R package for model-based clustering of mixed data (continuous, binary, integer). In this package, the model, submitted for publication in 2014, takes into account the main conditional dependencies between variables through Gaussian copula. Mixcluster performs the model selection and provides the best model according to Bayesian approaches.

- Participants: Matthieu Marbac-Lourdelle, Christophe Biernacki and Vincent Vandewalle
- Contact: Matthieu Marbac-Lourdelle
- URL: https://r-forge.r-project.org/R/?group_id=1939

## 6.14. Mixmod

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

Mixmod is a free toolbox for data mining and statistical learning designed for large and high dimensional data sets. Mixmod provides reliable estimation algorithms and relevant model selection criteria.

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

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

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

## 6.15. 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). A new advanced model concerning functional data is also available in 2016. 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: Vincent Kubicki, Christophe Biernacki and Serge Iovleff
- Contact: Christophe Biernacki
- URL: https://massiccc.lille.inria.fr/#/

## 6.16. RankCluster

FUNCTIONAL DESCRIPTION

Rankcluster package for R proposes a clustering tool for ranking data. Multivariate and partial rankings can be also taken into account. Rankcluster now supports tied ranking data.

- Participants: Christophe Biernacki, Julien Jacques and Quentin Grimonprez
- Contact: Quentin Grimonprez
- URL: https://cran.r-project.org/web/packages/Rankcluster/index.html

## 6.17. STK++

Statistical ToolKit
KEYWORDS: Statistics - Linear algebra - Framework
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

## 6.18. clere

FUNCTIONAL DESCRIPTION

The clere package for R proposes variable clustering in high dimensional linear regression. Available on CRAN and now published to an international journal dedicated to software: [24].

- Participants: Loïc Yengo, Christophe Biernacki and Julien Jacques
- Contact: Loïc Yengo
- URL: https://cran.r-project.org/web/packages/clere/index.html

## 6.19. 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

# 7. New Results

## 7.1. An oracle inequality for Quasi-Bayesian Non-Negative Matrix Factorisation

**Participant:** Benjamin Guedj.

We have extended the quasi-Bayesian perspective to the popular setting of non-negative matrix factorisation. This is a pivotal problem in machine learning (image segmentation, recommendation systems, audio source separation, ...) and we were able to propose an original estimator of the unobserved matrix. An oracle inequality is derived, along with several possible implementations. This work is now submitted to an international journal [38].

Joint work with Pierre Alquier.

## 7.2. PAC-Bayesian Online Clustering

**Participants:** Benjamin Guedj, Le Li.

We have extended the PAC-Bayesian framework to online learning. Our algorithm (called PACBO) performs online clustering of random sequences, and is supported by strong theoretical (regret bounds) and algorithmic (ergodicity of an MCMC implementation) results. This work is now submitted to an international journal [46].

Joint work with Sébastien Loustau.

## 7.3. Simpler PAC-Bayesian Bounds for Hostile Data

**Participant:** Benjamin Guedj.

We have introduced an original and much simpler way of deriving PAC-Bayesian bounds, through the use of $f$-divergences (therefore generalizing earlier works on Renyi's divergence and Kullback-Leibler divergence). This work is now submitted to an international conference [39].

Joint work with Pierre Alquier.

## 7.4. Clustering categorical functional data: Application to medical discharge letters

**Participants:** Cristian Preda, Cristina Preda, Vincent Vandewalle.

Categorical functional data represented by paths of a stochastic jump process are considered for clustering. For paths of the same length, the extension of the multiple correspondence analysis allows the use of well-known methods for clustering finite dimensional data. When the paths are of different lengths, the analysis is more complex. In this case, for Markov models we have proposed an EM algorithm to estimate a mixture of Markov processes. This work has been presented in a workshop [48].

## 7.5. Simultaneous dimension reduction and multi-objective clustering

**Participant:** Vincent Vandewalle.

In model based clustering of quantitative data it is often supposed that only one clustering variable explains the heterogeneity of all the others variables. However, when variables come from different sources, it is often unrealistic to suppose that the heterogeneity of the data can only be explained by one variable. If such an assumption is made, this could lead to a high number of clusters which could be difficult to interpret. A model based multi-objective clustering is proposed, it assumes the existence of several latent clustering variables, each one explaining the heterogeneity of the data on some clustering projection. In order to estimate the parameters of the model an EM algorithm is proposed, it mainly relies on a reinterpretation of the standard factorial discriminant analysis in a probabilistic way. The obtained results are projections of the data on some principal clustering components allowing some synthetic interpretation of the principal clusters raised by the data. This work has been presented in a conference [49].

## 7.6. Spatial Prediction of solar energy

**Participant:** Sophie Dabo.

Sophie Dabo-Niang's new result concern a work on spatial prediction of solar Energy in collaboration with some physicians and is now published [15].

This paper introduces a new approach for the forecasting of solar radiation series at a located station for very short time scale. We built a multivariate model in using few stations (3 stations). The proposed model is a spatio temporal vector autoregressive VAR model specifically designed for the analysis of spatially sparse spatio-temporal data. This model differs from classic linear models in using spatial and temporal parameters where the available predictors are the lagged values at each station. A spatial structure of stations is defined by the sequential introduction of predictors in the model. Moreover, an iterative strategy in the process of our model will select the necessary stations removing the uninteresting predictors and also selecting the optimal p-order. We studied the performance of this model. The metric error, the relative root mean squared error (rRMSE), is presented at different short time scales. Moreover, we compared the results of our model to simple and well known persistence model and those found in literature.

## 7.7. Multiple change-point detection

**Participants:** Alain Celisse, Guillemette Marot.

This is a joint work with Morgane Pierre-Jean and Guillem Rigaill (Univ. Evry).

The paper related to the work described in previous MODAL team reports (sections Kernel change point) has been pursuied and made available on Arxiv [42]. For recall, this work focuses on the problem of detecting abrupt changes arising in the full distribution of the observations (not only in the mean or variance). It provides greatly improved algorithms in terms of computational complexity (both in time and space). The computational and statistical performances of these new algorithms have been assessed through empirical experiments, which are detailed in the preprint.

## 7.8. Differential gene expression analysis

**Participants:** Alain Celisse, Guillemette Marot.

The use of empirical Bayesian techniques implemented in the R package metaMA has enabled to better understand Waldenstrom's macroglobulinemia. The new findings in Biology have been published in [18].

## 7.9. New concentration inequalities for the leave-$p$-out CV estimator

**Participant:** Alain Celisse.

New concentration inequalities have been established for the leave-$p$-out cross-validation estimator applied to assess the performance the $k$-nearest neighbour binary classifier. Joint work with Tristan Mary-Huard.

## 7.10. A new notion of stability for learning algorithms

**Participants:** Alain Celisse, Benjamin Guedj.

We introduced a new notion of stability for learning algorithms, which bridges the gap between the earlier uniform and hypothesis stability notions. It allows us to derive new PAC exponential concentration inequalities that apply to the Ridge regression algorithm as a first step. The first version of this work is presented in the preprint [41] and is now an active line of research.

## 7.11. Model for conditionally correlated categorical data

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

It is a model-based clustering proposal (called CMM for Conditional Modes Model) where categorical data are grouped into conditionally independent blocks. The corresponding block distribution is a parsimonious multinomial distribution where the few free parameters correspond to the most likely modality crossings, while the remaining probability mass is uniformly spread over the other modality crossings. The exact computation of the integrated complete-data likelihood allows to perform the model selection, by a Gibbs sampler, reducing the computing time consuming by parameter estimation and avoiding BIC criterion biases pointed out by our experiments. This work is now published in the international journal Advances in Data Analysis and Classification (Marbac et al, 2016). Furthermore, an R package (CoModes) is available on Rforge.

## 7.12. Mixture model for mixed kind of data

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

A mixture model of Gaussian copula allows to cluster mixed kind of data. Each component is composed by classical margins while the conditional dependencies between the variables is modeled by a Gaussian copula. The parameter estimation is performed by a Gibbs sampler. This work has been now accepted to an international journal [21]. Furthermore, an R package (MixCluster) is available on Rforge.

## 7.13. Degeneracy in multivariate Gaussian mixtures (complete data case)

**Participant:** Christophe Biernacki.

In the case of Gaussian mixtures, unbounded likelihood is an important theoretical and practical problem. Using the weak information that the latent sample size of each component has to be greater than the space dimension, a simple non-asymptotic stochastic lower bound on variances is derived. It is proved also that maximizing the likelihood under this data-driven constraint leads to consistent estimates. This work has been presented as an invited talk to the international workshop [28] and a paper for an international journal is been prepared.

This is a joined work with Gwënaelle Castellan of University of Lille.

## 7.14. Degeneracy in multivariate Gaussian mixtures (missing data case)

**Participants:** Christophe Biernacki, Vincent Vandewalle.

In the case of multivariate Gaussian mixtures, unbounded likelihood is an important theoretical and practical problem. However, in the case of missing data situations, this drawback is exacerbated for too reasons. Firstly, degeneracy frequence increases with missing data occurrence. Secondly, the EM dynamic is hardly detected since it implies linear grows of the log-likelihood, contrary to exponential grows in the complete data case, leading to computation waste and also high risk of erroneous estimates. Using the weak information that the latent sample size of each component (restricted to complete data) has to be greater than the space dimension, it is derived a simple contraint EM algorithm variant allowing to solve simultaneously both problems. This work has been presented to the international workshop [28] and a paper for an international journal is been prepared.

## 7.15. Data units selection in statistics

**Participant:** Christophe Biernacki.

Usually, the data unit definition is fixed by the practitioner but it can happen that it 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, categorical, counting, ...). It has led to an invitation to an international workshop [29] and a preprint is being to be prepared.

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

## 7.16. Label switching in Bayesian mixture model estimation

**Participants:** Christophe Biernacki, Benjamin Guedj, Vincent Vandewalle.

In the case of mixtures of distributions, it is well-known that the Bayesian posterior distribution is invariant to label switching, it means invariant to any renumbering of components. Consequences are important, typically leading to unuseful estimates like the posterior mean. Many attempts exist to solve this problem but it is advocated in this work that such a quest should be unfruitful since it is a direct consequence of the label non-identifiability of mixtures themselves. The present work proposes an original way to manage the label switching problem based on the Gibbs algorithm dynamic. The basic idea is to control the label switching probability along Gibbs iterations, controlled by both the sample size and the component overlap. An early version of this work has been presented as an invited talk to the international workshop [28].

## 7.17. Trade-off computation time and accuracy

**Participants:** Christophe Biernacki, Maxime Brunin, Alain Celisse.

Most estimates practically arise from algorithmic processes aiming at optimizing some standard, but usually only asymptotically relevant, criteria. Thus, the quality of the resulting estimate is a function of both the iteration number and also the involved sample size. An important question is to design accurate estimates while saving computation time, and we address it in the simplified context of linear regression here. Fixing the sample size, we focus on estimating an early stopping time of a gradient descent estimation process aiming at maximizing the likelihood. It appears that the accuracy gain of such a stopping time increases with the number of covariates, indicating potential interest of the method in real situations involving many covariates. A first version of this work has been presented to an international conference [27], and a preprint is being in progress.

## 7.18. 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 $R^2$ with a simple reading of the approximation quality, and provides then a novel variant of dimensionality reduction.

This work is under revision in an international journal [37] and it has also been presented to an international conference [25].

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

## 7.19. 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 descriptor equal for both learning and test samples. Indeed, for answering to 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 [33], with international audience.

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

## 7.20. Real-time audio sources classification

**Participants:** Christophe Biernacki, Maxime Baelde.

Recent research on machine learning focuses on audio source identification in complex environments. They rely on extracting features from audio signals and use machine learning techniques to model the sound classes. However, such techniques are often not optimized for a real-time implementation and in multi-source conditions. It is proposed here a new real-time audio single-source classification method based on a dictionary of sound models (that can be extended to a multi-source setting). The sound spectrums are modeled with mixture models and form a dictionary. The classification is based on a comparison with all the elements of the dictionary by computing likelihoods and the best match is used as a result. It is found that this technique outperforms classic methods within a temporal horizon of 0.5s per decision (achieved 6% of errors on a database composed of 50 classes). Future works will focus on the multi-sources classification and reduce the computational load. This work has been accepted in 2016 to be presented in 2017 to an international conference in Signal Processing [32].

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

## 7.21. Model-Based Co-clustering for Ordinal Data

**Participants:** Christophe Biernacki, Julien Jacques.

A model-based coclustering algorithm for ordinal data is presented. This algorithm relies on the latent block model using the BOS model (Biernacki and Jacques, 2015, Stat. Comput.) for ordinal data and a SEM-Gibbs algorithm for inference. Numerical experiments on simulated data illustrate the efficiency of the inference strategy. This work has been presented to an international workshop [30] and also to a national conference with an international audience [35].

## 7.22. Computational and statistical trade-offs in change-point detection

**Participants:** Christophe Biernacki, Maxime Brunin, Alain Celisse.

The change-point detection problem aims to detect changes in the distribution of observations collected over the time between the instants 1,...,T in the offline context. These changes occur at some instants called change-points. Our method provides consistent estimates of the change-points obtained by the Kernel Binary Segmentation algorithm with stopping rule (KBS). Moreover, the proposed method has a lower complexity in time and in space than the Kernel Dynamic Programming (KDP). This work has been presented to a national conference with an international audience [34].

## 7.23. MixtComp software for full mixed data

**Participants:** Christophe Biernacki, Vincent Kubicki.

MixtComp (Mixture Computation) is an integration software from the MODAL team for model-based clustering of mixed data. Its computing core is written in C++ and is accessed through an R interface. Its architecture allows to easily and quickly integrate new univariate models (under the conditional independence assumption) as they are published. The first phase of development was the implementation of three basic models (Gaussian, Multinomial, Poisson) with the native management of partially observed data (including intervals). It now implements models related to ordinal data (2015), rank data (2015) and functional data (2016), still with missing or partially missing data. The code is developed internally, and has been field-tested through several contracted partnerships (see the section about contracts). It is now referenced in the BIL database and the APP. It is available through a new web interface, called MASSICCC at https://massiccc.lille.inria.fr/#/ (see also the dedicated section). MixtComp has been presented to an invited talk in October 2016 at the Academy of Sciences in Tunisia [26].

## 7.24. MASSICCC platform for SaaS software availability

**Participants:** Christophe Biernacki, Vincent Kubicki, Matthieu Marbac Lourdelle.

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, it is planned to integrate the BlockCluster package. The MASSICCC platform gardually replaces the former BigStat platform available here: https://modal-research.lille.inria.fr/BigStat/. BigStat and MASSICCC have been both presented to an invited talk in October 2016 at the Academy of Sciences in Tunisia [26].

MASSICCC has led to a first short meeting in April 2016 in Lille for obtaining a feedback from company and academic users. Here is the link towards this event: Link. A second similar event is planned in February 2017 in Paris. Joint work with Jonas Renault and Josselin Demont (both at InriaTech).

The MASSICCC platform is available on https://massiccc.lille.inria.fr

## 7.25. CoModes package for correlated categorical variables

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

CoModes is an R package for model-based clustering of categorical data. In this package, the Conditional Modes Model (CMM), published in 2016 (Marbac et al, 2016), takes into account the main conditional dependencies between variables through particular modality crossings (so-called modes). CoModes performs the model selection and provides the best model according to the exact integrated likelihood criterion and the maximum likelihood estimates. It is available online on Rforge (https://r-forge.r-project.org/R/?group_id=1809).

### 7.26. MixCluster package for correlated mixed variables

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

MixCluster is an R package for model-based clustering of mixed data (continuous, binary, integer). In this package, the model, accepted for publication in 2016 [21], takes into account the main conditional dependencies between variables through Gaussian copula. Mixcluster performs the model selection and provides the best model according to Bayesian approaches. It is available online on Rforge (https://r-forge.r-project.org/R/?group_id=1939).

## 8. Bilateral Contracts and Grants with Industry

### 8.1. Arcelor-Mittal

**Participant:** Christophe Biernacki.

Arcelor-Mittal is a leader company in steel industry. This 11 months contract aims at optimizing predictive maintenance from mixed data (continuous, categorical, functional) provided by multiple sensors disseminated in steel production lines.

It is a joint work with Martin Bue and Vincent Kubicki (InriaTech engineers).

### 8.2. Banque Accord

**Participants:** Christophe Biernacki, Vincent Vandewalle.

Banque Accord is a credit scoring company. This 3 months contract aims at improving credit scoring performance by using the clustering principle inside the predictive process. In addition, directly managing mixed data (continuous, categorical, missing) has to be taken into account.

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

### 8.3. Vallourec

**Participant:** Christophe Biernacki.

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

It is a joint work with Vincent Kubicki (InriaTech engineer).

### 8.4. Cylande

**Participants:** Christophe Biernacki, Vincent Vandewalle.

Cylande is a software editor for retail. This 12 months contract aims at predicting future sales from past sales, including also many other available information.

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

### 8.5. NFID

**Participants:** Benjamin Guedj, Quentin Grimonprez.

NFID is the agency dedicated to innovation policies of the Hauts-de-France region.

This 3 months contract aims at clustering companies from Hauts-de-France based on their economic, social, environmental, innovation, activities data. The proposed methodology relies on the MixtComp software developed within Modal, and allows for the creation of a predictive analysis tool for NFID. This predictive tool aims at identifying regional companies with the highest innovative abilities, and has a great economic and politic impact.

# 9. Partnerships and Cooperations

## 9.1. Regional Initiatives

### *9.1.1. L'impact de l'évolution de l'état émotionnel et cognitif ressenti sur la reprise de l'activité de femmes atteintes d'un cancer du sein (Protocole FACEBROK)*

**Participant:** Sophie Dabo.

Partners: LAPCOS (EA 7278), UMR 9193 SCALab, LEM UMR 9221 LEM, Modal-Inria, EA CRDP

### *9.1.2. Main partners of bilille*

**Participant:** Guillemette Marot.

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

EA2694 (Univ. Lille 2, CHRU, Inria)

FRABIO, FR3688 (Univ. Lille 1,CNRS)

CBP / GFS (Univ. Lille 2, CHRU)

TAG (Univ. Lille 2, CNRS, INSERM, Institut Pasteur de Lille)

U1167 (Univ. Lille 2, CHRU, INSERM et Institut Pasteur de Lille)

U1011 (Univ. Lille 2, INSERM)

UMR8198 (Univ. Lille 1, CNRS)

LIGAN PM (Univ. Lille 2, CNRS)

BONSAI (Inria, Univ. Lille 1, CNRS)

Those teams are thus the main partners of MODAL concerning biostatistics for bioinformatics. Guillemette Marot is the leader 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 (deputy head of bilille)

P. Touzet, UMR8198 (deputy head of bilille)

V. Chouraki, U1167

M. Figeac, CBP / GFS

D. Hot, TAG

V. Leclère, Insitut Charles Viollette

M. Lensink, UGSF

### *9.1.3. New collaborations of the year linked to bilille, the bioinformatics and bioanalysis platform*

**Participants:** Guillemette Marot, Samuel Blanck.

Guillemette Marot has supervised the data analysis part or support in biostatistics tools testing for the following research projects involving Samuel Blanck or 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):

U 1011, H. Duez, circadiomics project

CIIL, J.C. Sirard, Flagnew project

JPARC, M.H. David, biostatistics related to DNAse-seq

### *9.1.4. Collaboration linked to SIRIC Oncolille*

**Participants:** Sophie Dabo, Guillemette Marot.

During the 'Plan Cancer 2' period, eight SIRICs ('Site de Recherche Intégrée sur le Cancer') were created in France, including the SIRIC ONCOLille (Link). More recently, the SFR Cancer has been created and Sophie Dabo-Niang is a member of the board that aims to create an Interdisciplinary Cancer Research Institute in Lille, based on ONCOLille. Guillemette Marot is still involved in several collaborations linked to cancer, through the projects analysed by the bilille platform.

## 9.2. National Initiatives

### *9.2.1. Programme of Investments for the Future (PIA)*

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

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

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

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

### *9.2.2. 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.

### *9.2.3. ANR*

**Participant:** Cristian Preda.

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. Other initiatives*

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.

## 9.3. European Initiatives

### *9.3.1. European Research Council*

Benjamin Guedj has participated in the 2017 Starting call of the European Research Council (ERC), by submitting a project (called BEAGLE, standing for PAC-Bayesian Agnostic Learning) in October 2016.

### *9.3.2. Collaborations with Major European Organizations*

EMS (European Mathematical Society), Sophie Dabo-Niang

Nominated (November 2016) as member of EMS-CDC (Committee of Developing counties)

CIMPA (International Center of Pure and Applied Mathematics), Sophie Dabo-Niang

Nominated (June 2016) as member

## 9.4. International Initiatives

### *9.4.1. Inria International Labs*

Sophie Dabo-Niang is a member of SIMERGE, a LIRIMA project-team started in January 2015. It includes researchers from Mistis (Inria Grenoble - Rhône-Alpes, France), LERSTAD (Laboratoire d'Etudes et de Recherches en Statistiques et Développement, Université Gaston Berger, Sénégal), IRD (Institut de Recherche pour le Développement, Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, Dakar, Sénégal) and LEM lab (Lille Economie et Management, Université Lille 1, 2, 3, Modal, Inria Lille Nord-Europe, France).

### *9.4.2. Inria Associate Teams Not Involved in an Inria International Labs*

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

*9.4.2.1. Informal International Partners*

Benjamin Guedj regularly collaborates with Olivier Wintenberger from Københavns Universitet (KU, Denmark).

Benjamin Guedj regularly collaborates with Sylvain Robbiano from University College London (UCL, UK).

## 9.5. International Research Visitors

### *9.5.1. Visits of International Scientists*

*9.5.1.1. Internships*

Rohit Uttam Bhagwat

Date: June 2016 - July 2016

Institution: Indian Institute of Science Education and Research, Kolkata (India)

Supervisor: Vincent Vandewalle

Siddharth Sharma Siddharth

Date: Nov 2015 - May 2016

Institution: LNM Institute of Information Technology (India)

Supervisor: Guillemette Marot

Miguel Assuncao
- Date: September 2016 - February 2017
- Institution: University of Lille
- Supervisor: Christophe Biernacki and Vincent Kubicki

Ghazouani Yannis
- Date: Oct 2015 - Sept 2016
- Institution: École Centrale Lille - VEKIA
- Supervisor: Alain Celisse

Hamza Tajmouati
- Date: Oct 2015 - Sept 2016
- Institution: École Centrale Lille
- Supervisor: Alain Celisse

Astha Gupta
- Date: May 2016 - Jul 2016
- Institution: BITS Pilani (India)
- Supervisor: Benjamin Guedj

Bhargav Srinivasa Desikan
- Date: Aug 2016 - Jul 2017
- Institution: BITS Pilani (India)
- Supervisor: Benjamin Guedj

### *9.5.2. Visits to International Teams*

*9.5.2.1. Research Stays Abroad*

Sophie Dabo-Niang has visited AIMS-Senegal (African Institute of Mathematical Sciences) and SIMERGE (Inria International Lab of University Gaston-Berger, Senegal) from July to mid-August, 2016.

# 10. Dissemination

## 10.1. Promoting Scientific Activities

### *10.1.1. Scientific Events Organisation*

*10.1.1.1. General Chair, Scientific Chair*

Christophe Biernacki co-organised a one-day meeting "Introduction aux modèles statistiques scalables: Modélisation, nouveaux paradigmes, ecosystème Big Data" on December 5th 2016 at IHP (Paris). The website is here https://bigdata-stat.sciencesconf.org/ and it brought together about 40 registered people (maximum number for this session).

Christophe Biernacki, Benjamin Guedj and Sophie Dabo-Niang were in the organizing committee of the workshop "Big Data: Modelling, Estimation and Selection" at École Centrale de Lille on June 9th and 10th 2016 (https://indico.math.cnrs.fr/event/830/).

Alain Celisse and Guillemette Marot co-organized SMPGD in Lille: The Statistical Methods for Post Genomic Data workshop is an annual meeting dedicated to statistical methods for post genomic data analysis. The aim of the workshop is to present works from mathematical to applied Statistics, but also new areas in high throughput Biology that could need new statistical developments. The workshop is usually organized around 3 to 4 invited speakers, and 3 to 4 invited sessions, and one session of contributed abstracts (oral presentations and posters). As SMPGD is especially interested in the statistical, mathematical, algorithmics or modelling questions raised by modern biology, presentations are expected to focus on these points.

Benjamin Guedj co-founded and co-organised a one-day workshop called Young Statisticians and Probabilists (YSP), in Paris in January 2016. The topics were sequential learning, random trees, random maps and random matrices theory. Nearly 80 PhD students, postdocs and young researchers attended.

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

*10.1.1.2. Member of the Organizing Committees*

Benjamin Guedj has been a member of the steering committee for the FEM (Forum Emploi Maths) in Paris in December 2016. The FEM is the largest mathematics jobs fair in France and gathers universities, students, graduates, companies and institutions. Over 2,000 people attended this edition.

Sophie Dabo-Niang has been a member of the steering committees for the following events.

- CIMPA Research School: "Statistical methods for evaluation of extreme risks": April, 5-15, 2016, St-Louis, Senegal
- Worksop: "Financial and actuarial Mathematics": July 11-15, 2016, AIMS-Mbour, Senegal
- The first AWMA (African Women in Mathematics Association) regional Forum, July 8-9, 2016, AIMS-Mbour, Senegal
- Session "EO075: Quantile regression models for dependent data ", "The 9th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2016)" (Link), December 9-11, 2016, Seville, Spain.
- Session "Asymptotic properties in nonparametric spatial problems ", "Third conference of the International Society for NonParametric Statistics (ISNPS)", June, 11-16, 2016, Avignon, France.
- Workshop "Statistical methods for recurrent data ", November 7th, 2016, Link.
- Workshop, "Learning with functional data", October, 7th, 2016, Lille, France. Link

Cristian Preda has co-organized the International Workshop on Applied Probability 2016 (Link).

Vincent Vandewalle is a member of the animation team of the bilille platform (https://wikis.univ-lille1.fr/bilille/animation). He has co-organized two scientific days, one in June 2016 on metagenomic analysis and another one in November 2016 on systems biology.

Serge Iovleff, Cristian Preda and Vincent Vandewalle have organised a one day workshop at Lille in October 2016 about learning with functional data. During this workshop, a large scope of methods for learning with functional data with application to various domains has been presented (https://functional-data.univ-lille1.fr).

### *10.1.2. Scientific Events Selection*

*10.1.2.1. Member of the Conference Program Committees*

Sophie Dabo-Niang has been a member of the scientific committees of two conferences:

- STAHY 2016 Workshop, September 26-27, Quebec, Canada.
- International colloquium on financial Econometrics, November, 18 -19, 2016, Rabat, Maroc. Link

*10.1.2.2. Reviewer*

Alain Celisse has acted as a reviewer for AISTATS 2016.

Benjamin Guedj has acted as a reviewer for NIPS 2016 and AISTATS 2017.

### *10.1.3. Journal*

*10.1.3.1. Member of the Editorial Boards*

Cristian Preda is a member of the editorial boards of

- Methodology and Computing in Applied Probability (Associate Editor)
- Romanian Journal of Mathematics and Computer Science" (Associate Editor)

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

Sophie Dabo-Niang is a member of the editorial board of *Revista Colombiana de Estadística*.

*10.1.3.2. Reviewer - Reviewing Activities*

The Modal team is involved in 34 scientific outlets, among which are the most prestigious international journals related to statistics.

1. Advances in Data Analysis and Classification (Christophe Biernacki)
2. Annals of Applied Statistics (Sophie Dabo-Niang)
3. Annals of Statistics (Alain Celisse)
4. Bernoulli (Alain Celisse, Cristian Preda)
5. BMC Research Notes (Christophe Biernacki)
6. BMC Medical Research Methodology (Benjamin Guedj)
7. Canadian Journal of Statistics (Christophe Biernacki)
8. Chemometrics (Cristian Preda)
9. Computational Statistics and Data Analysis (Christophe Biernacki, Sophie Dabo-Niang, Cristian Preda, Vincent Vandewalle)
10. Data Mining and Knowledge Discovery (Christophe Biernacki)
11. Electronic Journal of Statistics (Alain Celisse, Sophie Dabo-Niang)
12. ESAIM: Probability and Statistics (Sophie Dabo-Niang)
13. Journal de la SFdS (Christophe Biernacki)
14. Journal of Classification (Christophe Biernacki)
15. Journal of Computational and Graphical Statistics (Vincent Vandewalle)
16. Journal of Machine Learning Research (Christophe Biernacki, Alain Celisse)
17. Journal of Multivariate Analysis (Sophie Dabo-Niang, Benjamin Guedj)
18. Journal of Nonparametric statistics (Sophie Dabo-Niang)
19. Journal of Statistical Planning and Inference (Christophe Biernacki)
20. Journal of Statistical Software (Christophe Biernacki)
21. Journal of the American Statistical Association (Benjamin Guedj)
22. Journal of the Royal Statistical Society, series A (Benjamin Guedj)
23. Knowledge and Information Systems (Christophe Biernacki)
24. Mathematical Reviews (Benjamin Guedj)
25. Methodology and Computing in Applied Probability (Cristian Preda)
26. Metrika (Sophie Dabo-Niang)
27. Molecular Ecology Resources (Benjamin Guedj)
28. Neurocomputing (Benjamin Guedj)
29. Statistical Inference for Stochastic Processes (Sophie Dabo-Niang)
30. Statistical Methods and Applications (Sophie Dabo-Niang)
31. Statistics (Sophie Dabo-Niang)
32. Statistics and Computing (Serge Iovleff)
33. Statistics and Probability Letters (Sophie Dabo-Niang, Benjamin Guedj)
34. The American Statistician (Christophe Biernacki)

### 10.1.4. Invited Talks

Christophe Biernacki's talks in 2016:

- Working Group on Model-Based Clustering Summer Session, Paris, July 17-23, 2016, https://maths.ucd.ie/~brendan/wgmbc2016.html, [28]
- Workshop on Model-based Clustering and Classification, September 5-7, 2016, Catania (Italy), http://mbc2.unict.it/, [29]
- Académie des Sciences, des Lettres et des Arts, Journée Scientifique "Big Data & Data Science", October 28th 2016, Tunis (Tunisia), http://www.beitalhikma.tn/p7536/, [26]
- 9th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2016, ERCIM 2016), 9-11 December 2016, University of Seville, Spain http://cmstatistics.org/CMStatistics2016/, [27]
- Talk to the seminar of the "Laboratoire de Mathématiques de Besançon", February 29th 2016

Sophie Dabo-Niang's talks in 2016:

- Environmental Econometrics Day, Spatial Risk estimation and Application to environmental data, April, 24, 2016, Rabat, Marocco.
- CRoNoS FDA, satellite workshop of Compstat2016 (Link), Functional Binary Choice Models With Choice-Based Sampling, August, 26-28, Oviedo, Spain.
- AAS/AMU symposium on "Current Research Trends in Mathematical Sciences and applications", May, 17-20, 2016, UNESCO Chair of Mathematics, National Mathematical Centre (NMC), Abuja Nigeria.
- Learning with functional data, Functional Binary Choice Models With Choice-Based Sampling, October, 7, 2016, Lille, France.

Benjamin Guedj has been invited to deliver a talk to the 48èmes Journées de Statistique (JdS) of the French Statistical Society (June 2016, Montpellier, France).

Serge Iovleff gave a lightning talk entitled "MixAll: Un logiciel de classification non-supervisée" to the 5ème Rencontres R 2016.

Cristian Preda's talks in 2016:

1. International Workshop on Applied Probability (IWAP2016), 20-23 June, 2016, Toronto, Canada
2. 48e Journées de Statistique de la Société Francaise de Statistique, Montpellier, June 2016
3. 147th ICB Seminar Tenth International Seminar on statistics and clinical practice, May 15 - 18, 2016, Warsaw, Poland
4. 19-th Conference of the Romanian Society of Statistics and Probability, Universitatea Tehnica de Constructii Bucuresti, 27 mai 2016

### 10.1.5. Leadership within the Scientific Community

Christophe Biernacki is the president (since 2012) of the data mining and learning group of the French statistical association (SFdS, http://www.sfds.asso.fr/).

Benjamin Guedj is the president (since 2016) of the Young Statisticians group of the French Statistical Society.

Benjamin Guedj has joined the board of AMIES, the French Agency fostering collaborations between mathematicians and the private sector.

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

### 10.1.6. Scientific Expertise

Christophe Biernacki acted as an expert for two HCERES committees: one for teaching evaluation, one for research evaluation. He is also an elected member to the "Conseil National des Universités" (CNU) since October 2015.

Sophie Dabo-Niang offers expertise for Oreal's Award "Women in Science" since 2014.

### *10.1.7. Research Administration*

Christophe Biernacki was "Délégué Scientifique Adjoint" of the Inria Lille center until June 2016. He is still member of the "Bureau du Comité des Projets" (BCP) of the Inria Lille center.

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

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

Cristian Preda is a member of the Research Council of the University Lille 1.

## 10.2. Teaching - Supervision - Juries

### *10.2.1. Teaching*

Christophe Biernacki is the head of the M2 "Ingénierie Statistique et Numérique" http://mathematiques.univ-lille1.fr/Formation/ at University Lille 1.

Serge Iovleff is responsible for the Computer Science Licence semester.

Licence: Sophie Dabo-Niang, Probability, 24h, Lille 3, France

Licence: Serge Iovleff, Discrete Mathématics, 68H, Computer Science, IUT, Lille 1, France

Licence: Serge Iovleff, Linear Algebra 24H, Computer Science, IUT, Lille 1, France

Licence: Serge Iovleff, Analysis and Numerical Methods, 75H, Computer Science, IUT, Lille 1, France

Licence: Serge Iovleff, Graphs and languages, 36H, Computer Science, IUT, Lille 1, France

Licence: Serge Iovleff, Mathematical Modelisation, 38H, Computer Science, IUT, Lille 1, France

Licence: Serge Iovleff, Internships supervision, 15H, Computer Science, IUT, Lille 1, France

Licence: Serge Iovleff, Operational Research, 14H, Computer Science, IUT, Lille 1, France

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

Licence: Cristian Preda, Probability, 40h, L1, Polytech'Lille, France

Licence: Cristian Preda, Inferential Statistics, 50h, L1, Polytech'Lille, France

Licence: Vincent Vandewalle, Probability, 132h, L2, U. Lille 2, France

Licence: Vincent Vandewalle, Classification, 32h, L2, U. Lille 2, France

Licence: Vincent Vandewalle, Analysis, 24h, L2, U. Lille 2, France

Licence: Vincent Vandewalle, Machine Learning, 20h, L3, U. Lille 2, France

Master: Christophe Biernacki, coaching project, 10h, M1, U. Lille 1, France

Master: Christophe Biernacki, data analysis, 97.5h, M2, U. Lille 1, France

Master: Christophe Biernacki, coaching internship, 20h, M2, U. Lille 1, France

Master: Sophie Dabo-Niang, Advanced Statistics, 24h, Lille 3, France

Master: Sophie Dabo-Niang, Biostatistics Statistics, 40h, Master, Lille 1, France

Master: Sophie Dabo-Niang, Non-parametric Statistics, 24h, Master, UGB, Senegal

Master: Sophie Dabo-Niang, Spatial Statistics, 24h, Lille 3, France

Master: Benjamin Guedj, Statistical Learning: Theory and Algorithms, 18h, Lille 1, France

Master: Benjamin Guedj, Statistical Learning: Theory and Algorithms, 24h, Université Pierre & Marie Curie, France

Master: Benjamin Guedj, Statistical Learning: Theory and Algorithms, 30h, Institut de Statistique des Universités de Paris (ISUP), France

Master: Serge Iovleff, Object Oriented programming, 20H, Master Mathématiques Appliquées, Statistique - Ingénierie Mathématique, Lille 1, France

Master: Serge Iovleff, Probability and stochastic processes, AIMS (MBour-Sénégal)

Master: Guillemette Marot, Biostatistics, 44h, M1, U. Lille Droit et Santé, France

Master: Guillemette Marot, Coaching project, 12h, M1, U. Lille Droit et Santé, France

Master: Guillemette Marot, Supervised classification, 13h, M1, Polytech'Lille, France

Master: Cristian Preda, Data Analysis, 40h, M1, Polytech'Lille, France

Master: Cristian Preda, Biostatistics, 12h, M2, Polytech'Lille, France

Master: Vincent Vandewalle, Classification 60h, M1, U. Lille 1, France

Doctorat: Guillemette Marot, Data Analysis with R, 14h, U. Lille Droit et Santé, France

### *10.2.2. Supervision*

PhD: Jérémie Kellner, "Gaussian processes and kernel methods", Université Lille 1, 12/2016, Alain Celisse.

PhD: Quentin Grimonprez, "Variable selection in high dimensional setting with correlation", Inria DGA & Université Lille 1, 12/2016, Guillemette Marot, Julien Jacques, Alain Celisse.

PHD: Florence Loingeville, "Modèle linéaire généralisé hiérarchique Gamma-Poisson pour le contrôle de qualité en microbiologie", Université Lille 1, 01/2016, Cristian Preda.

PhD: Mohamed Yayaha, Lille 3, Sophie Dabo-Niang and Aboubacar Amiri.

PhD: Aladji Bassene, Lille 3 & UGB (Sénégal), Sophie Dabo-Niang.

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

PhD in progress: Maxime Baelde, "Identification, localisation, séparation temps réel de sources sonores dans les flux audio multi-canaux", A-Volute, Inria & Université Lille 1, 01/2016, Christophe Biernacki.

PhD in progress: Anne-Lise Bedenel, "Appariement de descripteurs évoluant dans le temps", PIXEO, Inria & Université Lille 1, 06/2015, Christophe Biernacki, Laetitia Jourdan.

PhD in progress: Adrien Ehrhardt, "Modèles prédictifs pour données volumineuses et biaisées. Application à l'amélioration du scoring en risque crédit", CACF, Inria & Université Lille 1, 06/2016, Christophe Biernacki, Philippe Heinrich, Vincent Vandewalle.

PhD in progress: Maxime Brunin, "Early stopping rules in statistical learning", 09/2014, Christophe Biernacki, Alain Celisse.

PhD in progress: Emad Drwesh, Lille 3, Sophie Dabo-Niang, Jérôme Foncel.

PhD in progress: Mohamed Salem Ahmed, Lille 3, Sophie Dabo-Niang and Mohamed Attouch.

PhD in progress: H. Sarter, "Outils statistiques pour la sélection de variables et l'intégration de données cliniques et omiques : développement et application au registre EPIMAD", 12/2016, C. Gower, Guillemette Marot.

### *10.2.3. Juries*

Christophe Biernacki participated as a reviewer to 5 PhD theses and 1 HdR committee, and as an examiner to 1 PhD thesis and 2 HdR committees. He also participated to 1 recruitment committee for a professor and was president of 1 recruitment committee for an assistant professor.

Alain Celisse has participated as an examiner to 1 PhD thesis.

Sophie Dabo-Niang has participated as an examiner to 4 PhD thesis.

Guillemette Marot was a member of two recruitment committees (MCU Univ. Nice, IE Univ. Lille). She was also an examiner to 1 PhD thesis.

Cristian Preda has participated as an examiner to 1 HdR committee.

Vincent Vandewalle has participated as an examiner to 1 PhD thesis.

## 10.3. Popularization

Christophe Biernacki has given about 10 talks during 2016 for institutions (Inria, universities, Ecole des Mines), companies and other related events. He gave also presentations towards students and industrials to the Xperium platform of the University of Lille about “Intelligence des données” (https://modal.lille.inria.fr/xperium/, http://learningcenters.nordpasdecalais.fr/innovation/fr/xperium). About 1,500 people came to this event during two years. He organized also a first short meeting in April 2016 in Lille for obtaining a feedback from company and academic users about the MASSICCC platform developped by the Modal and Select teams (https://massiccc.lille.inria.fr/#/). Here is the link towards this event: https://modal.lille.inria.fr/wikimodal/lib/exe/fetch.php?media=meeting_massiccc_7avril2016.pdf.

Sophie Dabo-Niang participates in the promotion of research among young children around a day of "Girls and Science, a light equation" organized in Lille (October 2016) and Senegal (Dakar, march 2016).

Benjamin Guedj has given a talk to high school students ("Terminale ISN") at Euratechnologies. The talk consisted in an overview of of machine learning impacts our everyday lives and how mathematicians contribute to learning in the big data era.

Vincent Vandewalle has given one presentation towards students to the Xperium platform of the University of Lille about “Intelligence des données” (https://modal.lille.inria.fr/xperium/, http://learningcenters.nordpasdecalais.fr/innovation/fr/xperium). He also has animated a formation on probabilities and statistics for middle School mathematics teachers through the Maison Pour la Science (http://www.maisons-pour-la-science.org/node/10641).

# 11. Bibliography

## Major publications by the team in recent years

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

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

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

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

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

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

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

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

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

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

## Publications of the year

### Articles in International Peer-Reviewed Journals

[11] A. AMIRI, S. DABO-NIANG, M. YAHAYA. *Non-parametric recursive density estimation for spatial data*, in "Comptes Rendus Mathématique", 2016 [*DOI :* 10.1016/J.CRMA.2015.10.010], https://hal.inria.fr/hal-01425935

[12] 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

[13] 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

[14] S. DABO-NIANG, S. GUILLAS, C. TERNYNCK. *More efficient kernel functional regression estimation with correlated errors*, in "Journal of Multivariate Analysis", 2016 [*DOI :* 10.1016/J.JMVA.2016.01.007], https://hal.inria.fr/hal-01425933

[15] S. DABO-NIANG, A. MAINA, T. SOUBDHAN, H. OULD-BABA. *Predictive spatio-temporal model for spatially sparse global solar radiation data*, in "Energy", June 2016, vol. 111, pp. 599–608 [*DOI :* 10.1016/J.ENERGY.2016.06.004], https://hal.inria.fr/hal-01425930

[16] 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

[17] A. DERMOUNE, C. PREDA. *Parametrizations, fixed and random effects*, in "Journal of Multivariate Analysis", November 2016 [*DOI :* 10.1016/J.JMVA.2016.11.001], https://hal.archives-ouvertes.fr/hal-01424782

[18] C. HERBAUX, E. BERTRAND, G. MAROT, C. ROUMIER, N. PORET, V. SOENEN, O. NIBOUREL, C. ROCHE-LESTIENNE, N. BROUCQSAULT, S. GALIÈGUE-ZOUITINA, E. BOYLE, G. FOUQUET, A. RENNEVILLE, S. TRICOT, F. MORSCHHAUSER, C. PREUDHOMME, B. QUESNEL, S. POULAIN, X. LELEU.

*BACH2 promotes indolent clinical presentation in Waldenström macroglobulinemia*, in "Oncotarget", 2016, https://hal.inria.fr/hal-01423307

[19] J. JACQUES, C. RUCKEBUSCH. *Model-based co-clustering for hyperspectral images*, in "Journal of Spectral Imaging", 2016, https://hal.archives-ouvertes.fr/hal-01367941

[20] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Latent class model with conditional dependency per modes to cluster categorical data*, in "Advances in Data Analysis and Classification", June 2016, vol. 10, n° 2, pp. 183–207 [*DOI :* 10.1007/s11634-016-0250-1], https://hal.archives-ouvertes.fr/hal-00950112

[21] 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

[22] P. MASSELOT, S. DABO-NIANG, F. CHEBANA, T. B. OUARDA. *Streamflow forecasting using functional regression*, in "Journal of Hydrology", April 2016, vol. 538, pp. 754–766 [*DOI :* 10.1016/J.JHYDROL.2016.04.048], https://hal.inria.fr/hal-01425931

[23] C. TERNYNCK, M. ALI BEN ALAYA, F. CHEBANA, S. DABO-NIANG, O. TAHA. *Streamflow hydrology classification using functional data analysis*, in "Journal of Hydrometeorology", 2016 [*DOI :* 10.1175/JHM-D-14-0200.1], https://hal.inria.fr/hal-01425934

[24] L. YENGO, J. JACQUES, C. BIERNACKI, M. CANOUIL. *Variable Clustering in High-Dimensional Linear Regression: The R Package clere*, in "The R Journal", 2016, vol. 8, n° 1, pp. 92-106, https://hal.archives-ouvertes.fr/hal-00940929

### Invited Conferences

[25] H. ALAWIEH, N. WICKER, C. BIERNACKI. *Projection under pairwise distance control*, in "9th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2016, ERCIM 2016)", Séville, Spain, December 2016, https://hal.archives-ouvertes.fr/hal-01420681

[26] C. BIERNACKI. *BigStat for Big Data: Big Data clustering through the BigStat SaaS platform*, in "Journée scientifique « Big Data & Data science »", Tunis, Tunisia, October 2016, https://hal.archives-ouvertes.fr/hal-01420650

[27] C. BIERNACKI, M. BRUNIN, A. CELISSE. *Computation time/accuracy trade-off and linear regression*, in "9th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2016, ERCIM 2016)", Séville, Spain, December 2016, https://hal.archives-ouvertes.fr/hal-01420659

[28] C. BIERNACKI, G. CASTELLAN, S. CHRETIEN, B. GUEDJ, V. VANDEWALLE. *Pitfalls in Mixtures from the Clustering Angle*, in "Working Group on Model-Based Clustering Summer Session", Paris, France, July 2016, https://hal.archives-ouvertes.fr/hal-01419755

[29] C. BIERNACKI, A. LOURME. *Unifying Data Units and Models in Statistics: Focus on (Co-)Clustering*, in "Workshop on Model-based Clustering and Classification (MBC2)", Catania, Italy, September 2016, https://hal.archives-ouvertes.fr/hal-01420657

[30] J. JACQUES, C. BIERNACKI. *Model-Based Co-clustering for Ordinal Data*, in "Working Group on Model-Based Clustering Summer Session", Paris, France, July 2016, https://hal.archives-ouvertes.fr/hal-01420648

[31] J. JACQUES, C. BIERNACKI. *Model-based co-clustering for ordinal data*, in "23th Summer Working Group on Model-Based Clustering of the Department of Statistics of the University of Washington", Paris, France, July 2016, https://hal.inria.fr/hal-01383912

### Conferences without Proceedings

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

[33] A.-L. BEDENEL, C. BIERNACKI, L. JOURDAN. *Appariement de descripteurs evoluant en temps Application a la comparaison d'assurance en ligne*, in "48èmes Journées de Statistique de la SFdS", Montpellier, France, May 2016, https://hal.archives-ouvertes.fr/hal-01420667

[34] M. BRUNIN, C. BIERNACKI, A. CELISSE. *Compromis précision-temps de calcul appliqué au problèeme de détection de ruptures*, in "48èmes Journées de Statistique de la SFdS", Montpellier, France, May 2016, https://hal.archives-ouvertes.fr/hal-01420669

[35] J. JACQUES, C. BIERNACKI. *Model-based co-clustering for ordinal data*, in "48èmes Journées de Statistique organisée par la Société Française de Statistique", Montpellier, France, 2016, https://hal.archives-ouvertes.fr/hal-01383927

[36] J. JACQUES, C. RUCKEBUSCH. *Co-clustering for hyperspectral images*, in "6th International Conference in Spectral Imaging", Chamonix, France, July 2016, https://hal.archives-ouvertes.fr/hal-01383918

### Other Publications

[37] H. ALAWIEH, N. WICKER, C. BIERNACKI. *Projection under pairwise distance controls*, December 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01420662

[38] P. ALQUIER, B. GUEDJ. *An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization*, August 2016, working paper or preprint, https://hal.inria.fr/hal-01251878

[39] P. ALQUIER, B. GUEDJ. *Simpler PAC-Bayesian Bounds for Hostile Data*, October 2016, working paper or preprint, https://hal.inria.fr/hal-01385064

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

[41] A. CELISSE, B. GUEDJ. *Stability revisited: new generalisation bounds for the Leave-one-Out*, August 2016, working paper or preprint, https://hal.inria.fr/hal-01355365

[42] A. CELISSE, G. MAROT, M. PIERRE-JEAN, G. RIGAILL. *New efficient algorithms for multiple change-point detection with kernels*, September 2016, working paper or preprint, https://hal.inria.fr/hal-01413230

[43] S. DABO-NIANG, M. S. AHMED. *Generalized Functional Linear Models under Choice-Based Sampling*, July 2016, working paper or preprint, https://hal.inria.fr/hal-01345918

[44] M. A. HASNAT, J. VELCIN, S. BONNEVAY, J. JACQUES. *Evolutionary clustering for categorical data using parametric links among multinomial mixture models*, March 2016, working paper or preprint, https://hal.inria.fr/hal-01204613

[45] L. LI, B. GUEDJ, S. LOUSTAU. *Clustering en ligne : le point de vue PAC-bayésien*, January 2016, working paper or preprint, https://hal.inria.fr/hal-01264934

[46] L. LI, B. GUEDJ, S. LOUSTAU. *PAC-Bayesian Online Clustering*, January 2016, working paper or preprint, https://hal.inria.fr/hal-01264233

[47] Y. B. SLIMEN, S. ALLIO, J. JACQUES. *Model-Based Co-clustering for Functional Data*, December 2016, working paper or preprint, https://hal.inria.fr/hal-01422756

[48] V. VANDEWALLE, C. PREDA. *Clustering categorical functional data Application to medical discharge letters Medical discharge letters*, July 2016, Working Group on Model-Based Clustering Summer Session: Paris, July 17-23, 2016, Poster, https://hal.inria.fr/hal-01424950

[49] V. VANDEWALLE. *Simultaneous dimension reduction and multi-objective clustering using probabilistic factorial discriminant analysis*, December 2016, CMStatistics 2016, https://hal.inria.fr/hal-01424965