IN PARTNERSHIP WITH:
CNRS

Université Lille 2

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2015

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, Researcher]

Faculty Members

Christophe Biernacki [Team leader, Univ. Lille I, Professor, HdR]
Alexandru Amarioarei [Univ. Lille I, until Oct 2015]
Alain Celisse [Univ. Lille I, Associate Professor]
Serge Iovleff [Univ. Lille I, Associate Professor]
Guillemette Marot [Univ. Lille II, Associate Professor]
Cristian Preda [Univ. Lille I, Professor, HdR]
Vincent Vandewalle [Univ. Lille II, Associate Professor]

Engineers

Samuel Blanck [Univ. Lille II]
Vincent Kubicki [Inria]

PhD Students

Anne Lise Bedenel [from Jun 2015, granted by CIFRE]
Maxime Brunin [Univ. Lille I]
Quentin Grimonprez [Inria]

Jérémie Kellner [Univ. Lille I]
Florence Loingeville [granted by CIFRE]
Clément Thery [until Feb 2015]

Post-Doctoral Fellow

Cristina Preda [Inria, from Mar 2015]

Administrative Assistants

Corinne Jamroz [Inria]
Anne Rejl [Inria]

Others

Cyrill Blache [Inria, Internship, until Feb 2015]
Siddharth Sharma [Inria, Internship, from Nov 2015]
Lamine Tikharoubine [Inria, Internship, from Apr 2015 until Aug 2015]
Simin Zeng [Inria, Internship, from Jul 2015 until Aug 2015]
Faicel Chamroukhi [Univ. Toulon, from Sep 2015]
Sophie Dabo [Univ. Lille III]
Olivier Delrieu
Julien Jacques [Univ. Lyon II, Associate Professor, HdR]

2. Overall Objectives

2.1. MOdel for Data Analysis and Learning

MODAL is a team focused on statistical methodology for data analysis (clustering, visualization) and learning (classification, density estimation). 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.

The team scientific objectives are split into two main 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 softwares 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 mixed data. It corresponds typically to continuous and categorical data but it includes also other widespread ones like ordinal, functional, rank,...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 mixed 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 of 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. Application domains

Potential application areas of statistical modeling for mixed data are (credit scoring, marketing, environment, medical, economic, hydrology,...), but MODAL favors applications related to biology, phylogeny, genetics and medicine. Members of the team are already experienced in these directions with complementary skills.

5. Highlights of the Year

5.1. Highlights of the Year

MODAL was implicated at the first level (general chair) in the organization of the main annual French conference in statistics gathering more than 400 participants (JdS 2015, see Section 10.1.1.1. General chair, scientific chair). It is the first time this conference is held in Lille since about 30 years.

MixtComp is the first package for clustering data with full mixed data (continuous, categorical, counting, ordinal, rank) with possibly missing or partially missing (intervals) data (see Section 6.15. MixtComp).

6. New Software and Platforms

6.1. BlockCluster

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), submitted for publication in 2014, 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: Christophe Biernacki
- 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

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

FUNCTIONAL DESCRIPTION

HDPenReg (High-Dimensional Penalized Regression) 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

KEYWORDS: Segmentation - Genomics - Marker selection - Biostatistics

SCIENTIFIC DESCRIPTION

MPAgenomics (Multi-Patient Analysis of Genomic markers) 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

FUNCTIONAL DESCRIPTION

MetaMA (Meta-analysis for MicroArrays) 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

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

FUNCTIONAL DESCRIPTION

MixAll (Clustering using Mixture Models) 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
- 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: Christophe Biernacki
- URL: https://r-forge.r-project.org/R/?group_id=1939

6.14. Mixmod

FUNCTIONAL DESCRIPTION

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

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

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

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

6.15. MixtComp

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: Vincent Kubicki, Christophe Biernacki and Serge Iovleff
- Contact: Christophe Biernacki
- URL: https://modal-research.lille.inria.fr/BigStat

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++

FUNCTIONAL DESCRIPTION

STK++ (C++ Statistical ToolKit) 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 submitted to an international journal dedicated to software.

- 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

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 (STK++ core library integration to R using Rcpp) 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. Functional data analysis applied to hydrological or environnemental data

Participant: Sophie Dabo.

The new results concern particularly functional data analysis applied to hydrological or environnemental data. First in a recent paper ([16]), two statistical techniques from the theory of functional data classification are adapted and applied for the analysis of flood hydrographs. Functional classification directly employs all data of a discharge time series and thus contains all available information on shape, peak, and timing. This potentially allows a better understanding and treatment of floods as well as other hydrological phenomena.

7.2. New functional regression model when data are auto-correlated

Participant: Sophie Dabo.

We develop a new functional regression model when data are auto-correlated, in collaboration with Serge Guillas (University of College London) and Camille Ternynck (University of Lille 2). This work will appear in Journal of Multivariate Analysis. (Dabo-Niang, S, Guillas, S et Ternynck, C. (2016). More efficient kernel functional spatial regression estimation with autocorrelated errors. *Journal of Multivariate Analysis*). In this work we introduce a new procedure for the estimation in the nonlinear functional regression model where the explanatory variable takes values in an abstract function space and the residual process is autocorrelated. The procedure consists in a pre-whitening transformation of the dependent variable based on the estimated autocorrelation. We establish both consistency and asymptotic normality of the regression function estimate. For kernel methods encountered in the literature, the correlation structure is commonly ignored (the so-called "working independence estimator"); we show here that there is a strong benefit in taking into account the autocorrelation in the error process. We also find that the improvement in efficiency can be large in our functional setting, up to 25% in the presence of high autocorrelation levels. Concerning spatial data, we develop a new spatial prediction method that takes into account the spatial dependence. This work will appear in Journal of Nonparametric Statistics (Dabo-Niang, Ternynck, C., Yao, A.-F. (2016). Nonparametric prediction in the multivariate spatial context. *Journal of Nonparametric Statistics*)

7.3. Differential gene expression analysis

Participant: Guillemette Marot.

The use of empirical Bayesian techniques implemented in the Bioconductor package `limma` has enabled to better understand Waldenstrom's macroglobulinemia. Gene Set enrichment analysis was also performed after differential analysis. The new findings in Biology have been published in [21].

7.4. Evolutionary clustering for categorical data

Participant: Julien Jacques.

This is a joint work with Md Abul Hasnat, Julien Velcin and Stephane Bonnevay (Univ. de Lyon).

An evolutionary clustering algorithm for categorical data has been developed, based on parametric links between multinomial mixture models. This model has been used to study the evolution of opinions in Twitter data. A Preprint of this work is available [54].

7.5. 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 conference [34].

7.6. Degeneracy in Gaussian Mixtures with missing data

Participants: Christophe Biernacki, Vincent Vandewalle.

The missing data problem is well-known for statisticians but its frequency increases with the growing size of modern datasets. In Gaussian model-based clustering, the EM algorithm easily takes into account such data by dealing with two kinds of latent levels: the components and the variables. However, the quite familiar degeneracy problem in Gaussian mixtures is aggravated during the EM runs. Indeed, numerical experiments clearly reveal that degeneracy is quite slow and also more frequent than with complete data. In practice, such situations are difficult to detect efficiently. Consequently, degenerated solutions may be confused with valuable solutions and, in addition, computing time may be wasted through wrong runs. A simple condition on the latent partition to avoid degeneracy has been exhibited, and a constrained version of the Stochastic EM (SEM) algorithm satisfying this condition has been proposed. This work has been presented in a conference [33].

7.7. Model for conditionally correlated categorical data

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

An extension of the latent class model is proposed for clustering categorical data by relaxing the classical class conditional independence assumption of variables. In this model (called CCM for Conditional Correlated Model), variables are grouped into inter-independent and intra-dependent blocks in order to consider the main intra-class correlations. The dependence between variables grouped into the same block is taken into account by mixing two extreme distributions, which are respectively the independence and the maximum dependence ones. In the conditionally correlated data case, this approach is expected to reduce biases involved by the latent class model and to produce a meaningful model with few additional parameters. The parameters estimation by maximum likelihood is performed by an EM algorithm while a MCMC algorithm avoiding combinatorial problems involved by the block structure search is used for model selection. Applications on sociological and biological data sets bring out the proposed model interest. These results strengthen the idea that the proposed model is meaningful and that biases induced by the conditional independence assumption of the latent class model are reduced. This work is published [20] . Furthermore, an R package (Clustericat) is available on Rforge(see https://github.com/rforge/clustericat).

7.8. Model-based clustering for multivariate partial ordinal data

Participants: Christophe Biernacki, Julien Jacques.

We design the first univariate probability distribution for ordinal data which strictly respects the ordinal nature of data. More precisely, it relies only on order comparisons between modalities, the proposed distribution being obtained by modeling the data generating process which is assumed, from optimality arguments, to be a stochastic binary search algorithm in a sorted table. The resulting distribution is natively governed by two meaningful parameters (position and precision) and has very appealing properties: decrease around the mode, shape tuning from uniformity to a Dirac, identifiability. Moreover, it is easily estimated by an EM algorithm since the path in the stochastic binary search algorithm is missing. Using then the classical latent class assumption, the previous univariate ordinal model is straightforwardly extended to model-based clustering for multivariate ordinal data. Again, parameters of this mixture model are estimated by an EM algorithm. Both simulated and real data sets illustrate the great potential of this model by its ability to parsimoniously identify particularly relevant clusters which were unsuspected by some traditional competitors. This work is now published in an international journal [12] and is also currently available in the MixtComp software at https://modal-research.lille.inria.fr/BigStat/

7.9. Semi-Linear Auto-Associative Model

Participant: Serge Iovleff.

We design a new model for data analysis which is a generalization of the probabilistic PCA. The interpretation properties of the PCA are preserved while presence of non-linear repartitions in data can be detected and adjusted using B-spline regression. This model has been published in [18].

8. Bilateral Contracts and Grants with Industry

8.1. Genoscreen

Participant: Guillemette Marot.

Subject: Genoscreen is a company which offers innovative solutions in genomics and molecular biology. New technologies such as high throughput sequencing have raised statistical questions to analyse metagenomic data. Formation and expertise has been provided to this company to help them analyse this new type of data.

8.2. Bilateral Contracts with Industry

Participant: Julien Jacques.

ORANGE Labs : contrat de recherche pour l’extraction de connaissances dans de gros volumes de données hétérogènes pour la gestion automatique des réseaux radio, en lien avec le financement de la thèse CIFRE de Yosra Ben Slimen.

8.3. Arcelor-Mittal

Participants: Christophe Biernacki, Clément Thery.

Subject: Supervised and semi-supervised classification on large data bases mixing qualitative and quantitative variables. Arcelor Mittal faced some quality problems in the steel production which lead to supervised and semisupervised classification involving (1) a small number of individuals comparing to the numbers of variables, (2) heterogeneous variables, typically categorical and continous variables and (3) potentially highly correlated variables. A PhD CIFRE grant started on May 2011 on this topic and has finished on July 8th 2015. It has led also the the CorReg package, available on the CRAN (https://cran.r-project.org/web/packages/CorReg/index.html) and referenced on the Inria BIL application

8.4. Auchan

Participants: Christophe Biernacki, Serge Iovleff, Vincent Vandewalle.

Subject: Groupe Auchan SA is a French international retail group and multinational corporation headquartered in Croix. It is one of the world's principal distribution groups with a presence in 12 countries and 269,000 employees. The aim of the two months contracts (It started late 2014 and finished early in 2015) between Auchan and Modal is to identify human factors which significantly impact the economical results of the company. From a scientific point of view, it corresponds to regression studies (simple and mixture regression) with missing data and correlated data.

8.5. PIXEO

Participants: Christophe Biernacki, Anne Lise Bedenel.

Subject: PIXEO is a company allowing online comparisons of insurances. A PhD thesis for optimizing the workflow related to this activity started in June 2015, with co-supervisision of Laetitia Jourdan of the Dolphin Inria team. The title of the thesis is "Supervised and unsupervised classification with descriptors evolving in time. Application to online comparisons of insurances." Before the beginning of the thesis, a preliminary contract has been established since October 2014 until May 2015, in order to prepare precisely the research subject. It was a work in collaboration with two members of the Dolphin Inria team (Laetitia Jourdan and Marie-Eléonore Marmion).

8.6. Cylande

Participants: Christophe Biernacki, Etienne Goffinet, Vincent Kubicki, Vincent Vandewalle.

Subject: Cylande is a company which provides software solutions for retail. The aim of the contract is to provide statistical tools for optimal management delivery. The proposed solution relies on density estimation and also on model-based clustering, both for mixed data (count data, categorical data, continuous data). It should involved the MixtComp software (referenced on the Inria BIL application) developped by the team. It is a 12 months contract, started on October 1st 2015.

9. Partnerships and Cooperations

9.1. Regional Initiatives

Christophe Biernacki has some contracts and/or PhD theses with regional companies: Arcelor-Mittal (thesis), Auchan (contract), PIXEO (contract and thesis), Cylande (contract).

9.1.1. Collaborations within PSo-Innov

Participant: Sophie Dabo.

Sophie Dabo is a member of the regional emergent project *Précarité, Solidarité, vers un accompagnement innovant des personnes en difficultés d'une association spécialisée* with the LGI2A, CRIL, Discontinuité, LEM, APSA-Pas-de-Calais and coordinator: Issam Nouaouri (issam.nouaouri@univ-artois.fr).

9.1.2. MPAGenomics2

Participants: Samuel Blanck, 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. This last one financed the project MPAGenomics2, coordinated by Guillemette Marot, to biologically validate on cohorts of patients suffering from leukaemia the tools developed by the Development Technological Action MPAGenomics. The project lasted five months and other partners were Functional Genomics platform from Univ. Lille 2, INSERM UMR-S 1172 and biology pathology center of Lille hospital.

9.2. National Initiatives

9.2.1. ANR ClinMine

Participants: Julien Jacques, Cristian Preda, Vincent Vandewalle.

Modal team is member of ClinMine ANR project (http://www.lifl.fr/ClinMine/pmwiki/index.php) in charge with statistical methdology. Collaborators : LIFL, CHRU Lille, CHU Montpellier, ALICANTE, GHICL.

9.2.2. ANR Imagiweb

Participant: Julien Jacques.

Julien Jacques is member of Imagiweb ANR project (http://mediamining.univ-lyon2.fr/people/velcin/imagiweb/) as member of the ERIC laboratory (Univ. de Lyon).

9.2.3. ANR Calibration

Participant: Alain Celisse.

Alain Celisse is a member of the Calibration ANR project (https://sites.google.com/site/anrcalibration/anr-calibration) in charge with statistical methdology. Collaborators : Select, ENS Cachan, Université Paris-Sud, Université Nice, Université Paul Sabatier de Toulouse.

9.2.4. Working groups

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

Sophie Dabo belongs to the working groups

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

Guillemette Marot belongs to the StatOmique working group

Julien Jacques belongs to the Working Group on Model Based Clustering (University of Washington)

Benjamin Guedj belongs to the following GdR of CNRS: ISIS (local referee for Inria Lille - Nord Europe), MaDICS, MASCOT-NUM (local referee for Inria Lille - Nord Europe).

Alain Celisse belongs to the Statistics for Systems Biology group (SSB) in Paris.

Alain Celisse belongs to a working group on change-point detection with people from Lancaster university (UK).

9.3. International Initiatives

9.3.1. SIMERGE

Participant: Sophie Dabo.

SIMERGE is 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

LEM lab, Lille Economie et Management, Université Lille 1, 2, 3

9.3.2. Inria International Partners

9.3.2.1. Informal International Partners

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

9.4. International Research Visitors

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

9.4.1. Visits of International Scientists

Sylvain Robbiano (March 2015 - University College London, UK) and Pierre Alquier (April 2015 - ENSAE ParisTech, France) have visited Benjamin Guedj. Those two visits have been followed by the submission of two research papers (Nov. 2015 and Jan. 2016, respectively).

9.4.1.1. Internships

Siddharth Sharma Siddharth

Date: Nov 2015 - May 2016

Institution: LNM Institute of Information Technology (India)

Supervisor: Guillemette Marot

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. General chair, scientific chair

Christophe Biernacki was the General chair of the “47èmes Journées de Statistique de la SFdS” in Lille (JdS 2015, http://jds2015.sfds.asso.fr/). It holds during one week from June 1st to 5th 2015 with about 200 talks and about 400 registrations. It is the main annual French statistical conference, with international audience.

Christophe Biernacki co-organised a one-day meeting “Big-Data : Une vision globale Gestion, Analyse, Ethique et Logiciels” in Paris on March 13 2015 (http://www.sfds.asso.fr/393-Big-Data). It brought together about 80 registered people.

Sophie Dabo participates to the organisation of the Ecole Mathématique Africaine, 30/11-5/12 2015, Institut Polytechnique de Yamoussoukro, Côte d’ivoire

Benjamin Guedj is a founder and co-organizer of the YSP seminar (Young Statisticians and Probabilists), from the SFdS (Société Française de Statistique).

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

Benjamin Guedj is the co-organizer of the "Ateliers de la Statistique" from the SFdS.

10.1.1.2. Member of the organizing committees

Christophe Biernacki co-organized a one-day meeting “Statistique et données massives : enjeux et perspectives” in Paris on October 13th 2015 (http://bigdata2015.sfds.asso.fr/comite-dorganisation/).

Sophie Dabo co-organisate session "Asymptotic properties in nonparametric problems (EO116)" in "The 8th International Conference of the ERCIM WG on Computational and Methodological Statistics" (CMStatistics 2015)

Sophie Dabo co-organisate of "Forum of Young Mathematicians", Lille, 27-29, November, 2015

Guillemette Marot is a member of the organizing committee of seminars from Bilille platform. More information about all seminars is available on https://wikis.univ-lille1.fr/bilille/animation.

Benjamin Guedj has been a member of the steering committee of the 47th Journées de Statistique (JdS 2015, Lille, June 2015) from the SFdS.

10.1.2. Scientific events selection

10.1.2.1. Member of the conference program committees

Christophe Biernacki was in the conference program committee of the "Sixièmes rencontres des jeunes statisticiens" in parc ornithologique du Teich (near Bordeaux) on August 28th to September 2nd (http://rencontres-jeunes-statisticiens.sfds.asso.fr/).

Julien Jacques is member of the program committees of StatLearn'15 (Grenoble, april 2015) and StatLearn'16 (Vannes, april 2016).

Benjamin Guedj has been a member of the program committee of CaP'2015 (Conférence Francophone sur l'Apprentissage automatique, Lille, July 2015).

Alain Celisse was a member of the Journées de Statistique de la SFdS program committe.

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).

Sophie Dabo is an Associate editor of *Revista Colombiana de Estadistica*

Benjamin Guedj has been a reviewer for the conference CaP'2015 (Conférence Francophone sur l'Apprentissage automatique, Lille, July 2015).

Cristian Preda is an Associate Editor for the Journal Methodology and Computing in Applied Probability (http://www.springer.com/statistics/journal/11009) and Romanian Journal of Mathematics and Computer Science (www.rjm-cs.ro/).

10.1.3.2. Reviewer - Reviewing activities

Christophe Biernacki has acted as a reviewer for the a dozen journal papers: *Journal de la Société française de Statistique (JSFdS)*, *Journal of Classification (JoC)*, *Computational Statistics and Data Analysis (CSDA)*, *Journal of Statistical and Planning Inference (JSPI)*, *Advances in Data Analysis and Classification (ADAC)*, *Communication in Statistics – Theory and Methods*, *Data Mining and Knowledge Discovery (DAMI)*, *Journal of Statistical Software (JSS)*, *Statistica Sinica*, *Electronical Journal of Statistics (EJS)*, *Canadian Journal of Statistics (CJS).*

Sophie Dabo is reviewer of *Statistical Inference for Stochastic Processes*, *Computational Statistics and Data Analysis*, *Statistics*, *Journal Afrika Statistika*, *Journal of Multivariate Analysis*, *Journal of Nonparametric statistics*, *Annales de l'ISUP*, *Electronic journal of statistics*, *Metika, Annals of Statistics*

Julien Jacques has referee in 2015 papers for CSDA and Statistical Papier.

Benjamin Guedj is a reviewer for the following journals: *Journal of the Royal Statistical Society (Series A)*, *Journal of the American Statistical Association*, *Molecular Ecology Resources*, *Journal of Multivariate Analysis*, *BMC Medical Research Methodology*, *Neurocomputing*.

Alain Celisse is a reviewer for the folloowing journals: *Annals of Statistics*, *JMLR*, *Bernoulli*, *JMVA*,...,

Guillemette Marot has reviewed in 2015 papers for BMC Bioinformatics and the Journal of Bioinformatics Research Studies.

10.1.4. Invited talks

Christophe Biernacki has been invited to give a talk at the following meetings:

- a tutorial “Model-based clustering/imputation with missing/binned/mixed data using the new software MixtComp” at the MissData conference in Rennes on June 18-19 2015 (http://missdata2015.agrocampus-ouest.fr/infoglueDeliverLive/).
- two lectures, named “Part I: CorReg Linear regression with correlated and numerous data” and “Part II: MixtComp Supervised classification with mixed data, missing data and uncertain data”, at the National Institute for Public Health (RIVM) in Utrecht (The Netherlands) on April 8-9 2015.
- a lecture “Clustering: evolution of methods to meet new challenges” to the one-day clustering meeting of Orange Labs, Issy Les Moulineaux on October 20th 2015 (http://www.vincentlemaire-labs.fr/Clustering2015/).
- a research talk at CMStatistics 2015 (ERCIM 2015) in London (UK) on December 12-14 2015 [25].

Sophie Dabo give an invited talk at *Tunisian Association of Statistics and Applications Conference*, march 2015 and an invited talk at the *African Women in Mathematics Association Conference*, july 2015, Kenya.

Julien Jacques has given an invited talks at CMStatistics 2015 (London, UK, december 2015), ISI2015 (Rio de Janeiro, Brasil, July 2015). He has been also invited at Università degli Studi di Napoli Federico II to give a talk on a workshop on Ordinal Data Modelling.

Alain Celisse has given invited talks at

- Nice: Calibration project meeting
- Rennes: talk at the seminar
- LSTA/UMPC: talk at the seminar
- IHP: talk at the séminaire parisien de statistique

Vincent Vandewalle has given invited talks at

- Classification society meeting, June 2015, Mc Master University, Hamilton, Canada [31]
- 8th International Conference of the ERCIM WG on Computational and Methodological Statistics, December 2015, Senate House, London, United Kingdown [33]

10.1.5. Scientific expertise

Christophe Biernacki is an elected member to the “Conseil National des Universités” (CNU) since October 2015.

Sophie Dabo, Expertise of l’Oreal’s Award "Womens in Science" since 2014.

Julien Jacques is referee for ANRT PhD grant.

10.1.6. Research administration

Christophe Biernacki is “Délégué Scientifique Adjoint" of the Inria Lille center since August 2014.

Julien Jacques is responsible of the Data Mining & Decision team of ERIC laboratory (Université de Lyon).

Vincent Vandewalle, with Chloé Friguet (Laboratoire de Mathématiques de Bretagne Atlantique) and Frédérique Letué (Laboratoire Jean Kuntzmann): organisation of a roundtable during the "47èmes Journées de Statistique de la SFdS" about how to teach the big data [51].

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

Julien Jacques is the coordinator of the European Master on Data Mining and Knowledge Management (Université de Lyon).
Vincent Vandewalle was head of the DUT STID at IUT of University Lille 2 until June 2015.

Master: Christophe Biernacki, mathematical statistics, 60h, M1, Lille 1, France

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

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

Master: Christophe Biernacki, analysis of variance and experimental design, 22.5h, M2, Lille 1, France

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

Master: Sophie Dabo, "Spatial Statistics" 24h, M2, Lille 3, France

Master: Sophie Dabo, "Advanced Statistics" 24h, M2, Lille 3, France

Master: Sophie Dabo, "Non-prametric Statistics" 25h, M2, UGB, Senegal

Licence: Sophie Dabo, "Probability" , 24h, L2, Lille 3, France

Licence: Quentin Grimonprez, TP Probabilités, 12h, L3, Polytech'Lille, France

Master: Quentin Grimonprez, TP Classification Supervisée, 32h, M1, Polytech'Lille, France

Master: Quentin Grimonprez, TP Séries Temporelles, 14h, M2, Polytech'Lille, France

Licence: G. Marot, Biostatistics, 9h, L1, U. Lille 2, France

Master: G. Marot, Biostatistics, 45h, M1, U. Lille 2, France

Doctorat: G. Marot, Data Analysis with R, 24.5h, U. Lille 2, France

Master: B. Guedj, Statistical Learning, 10h, ENSAE ParisTech, Paris, France. 20 students.

Master: B. Guedj, Statistical Learning, 10h, Université de Lille, master de mathématiques, M2, parcours mathématiques et finance. 8 students.

Licence: S. Iovleff, Analysis and numerical methods, 28h, L1, U. Lille 1, France.

Licence: S. Iovleff, Linear Algebra, 74h, L1, U. Lille 1, France

Licence: S. Iovleff, Operational research, 28h, L2, U. Lille 1,France

Licence: S. Iovleff, Probability and Statistics, 32h, L3, U. Lille 1, France.

Master: S. Iovleff, Monte Carlo method, 30h, M1, U. Lille 1, France

Master: Maxime Brunin, TP du module "Modèle linéaire", 12h, M1, Polytech Lille, FRANCE.

Licence: V. Vandewalle, Probability, 142h, L2, U. Lille 2, France

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

Licence: V. Vandewalle, Case study and survey, 52h, L2, U. Lille 2, France

Licence: V. Vandewalle, Analysis, 28h, L2, U. Lille 2, France

Master: V. Vandewalle, Classification 34h, M1, U. Lille 1, France

Licence: Alain Celisse, Proba-Stat, 64h, niveau L2, université Lille, France

Licence: Alain Celisse, Analyse numérique, 32h, niveau L1, université Lille, France

Formation continue: Alain Celisse, Analyse numérique, 24h, niveau L1, université Lille, France

Licence: Alain Celisse, Algèbre, 68h, niveau L2, université Lille, France

Master: Alain Celisse, mémoire de Proba-Stat, 10h, niveau M2, université Lille, France

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

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

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

Licence: Cristian Preda, Biostatistics, 12h, M2, Polytech'LIlle, France

Licence: Cristian Preda, Functional data analysis, 12h, M2, Lille 1, France

10.2.2. Supervision

PhD: Clément Thery, Model-based covariable decorrelation in linear regression (CorReg). Application to missing data and to steel industry, University Lille 1, defended on July 8th 2015, Christophe Biernacki.

PhD: Stéphane Bouka, Modélisation non-paramétrique pour des observations spatialement dépendantes, Lille 3 et USTM Dabon, 21/12/2015, Sophie Dabo, Guy-martial Nkiet

PhD in progress: Aladji Bassene (2011-), co-tutelle between Lille 3 (Sophie Dabo) and UGB (Sénégal) (defense, April 2016)

PhD in progress: Emad Drwesh (2012-)(defense, december 2016), Lille 3, Sophie Dabo, Jérôme Foncel

PhD in progress: Mohamed Yayaha (2012-) (defense, december 2016), Lille 3, Sophie Dabo and Aboubacar Amiri

PhD in progress: Mohamed Ould Yehdhih (2014-) (defense, december 2017), Lille 3, Sophie Dabo and Mohamed Attouch

PhD in progress: Komi Nagbe, Prévision de production et de consommation d'énergie renouvelable, septembre 2015, Julien Jacques

PhD in progress: Yosra Ben Slimen, Extraction de connaissances dans de gros volumes de données hétérogènes (Big Data) pour la gestion automatique des réseaux radio, juin 2015, Julien Jacques

PhD in progress : Florence Loingeville, Mise en place d'outils statistiques spécifiques au contrôle de procédé en analyse microbiologique, décembre 2012, Julien Jacques and Cristian Preda

PhD in progress : Le LI, "PAC-Bayesian Online Clustering: theory and algorithms", 01/11/2014, together with Sébastien Loustau (Université d'Angers).

PhD in progess: Jérémie Kellner, Processus gaussien dans les RKHS et test d'adéquation, Université de Lille, Alain Celisse and Christophe Biernacki

PhD in progress: Quentin Grimonprez, Sélection de variable en très grande dimension par prise en compte de la dépendance, Alain Celisse, Guillemette Marot and Julien Jacques

PhD in progress: Maxime Brunin, Compromis temps de calcul-précision statistique, Alain Celisse and Christophe Biernacki

PhD in progress: Anne-Lise Bedenel, supervised and unsupervised classification with descriptors evolving in time. Application to online comparisons of insurances, Université Lille 1, Christophe Biernacki and Lætitia Jourdan

10.2.3. Juries

- PhD: Sophie Dabo has been examinator for Hassan Maatouk, Ecole of Mines of St-Etienne, 1/10/2015
- PhD: Sophie Dabo has been examinator for Sobom Matthieu Somé, University Franche-Comté, November 16, 2015.
- PhD: Sophie Dabo has been examinator for Abdoulaye Faye, Université Blaise Pascal, Clermont-Ferrand, december, 8, 2015.
- PhD: Sophie Dabo has been examinator for Zahraa Salloum, University of Lyon 2, 19/1/2016
- PhD: Julien Jacques has been referee for the PhD thesis of Amaury Labenne (Université de Bordeaux) and Henri Wallard (CNAM Paris)
- Christophe Biernacki participated as a reviewer to 5 PhD theses and 1 HdR, and as an examinator to 1 PhD thesis and 1 HdR.
- Critian Preda has been for the PḧD thesis of Alin Rusu (Facultatea de Matematica, Universitatea Bucuresti, June 2015).

Guillemette Marot was a member of an INSERM competition jury (IE number 13 BAP A)

10.3. Popularization

- **Research Sensibilisation**
  - Christophe Biernacki has given a talk “Big Stat / MixtComp: A SaaS platform for easy Big Data analysis” to the one-day meeting Data For You at EuraTechnologie in Lille on November 26th 2015 (http://www.euratechnologies.com/actualites/2015/10/data-you-10381).
  - Sophie Dabo has participated to research valorisation days for high school students female.
  - Benjamin Guedj has participated in a meeting with undergraduate students at Université de Lille to promote research positions in mathematics (January 2015).
  - Benjamin Guedj has been a speaker for the “30 minutes de science” seminar (Inria Lille - Nord Europe, April 2015).
  - Vincent Vandewalle has given a talk “Clustering et modèles prédictif” to the third R&D du Plateau at EuraTechnologie in Lille on April 22th 2015 http://www.euratechnologies.com/actualites/2015/04/3eme-r-dv-plateau-10038
- **XPérium Lille 1**

  Participants: Christophe Biernacki, Maxime Brunin, Quentin Grimonprez, Vincent Kubicki, Vincent Vandewalle

  Vulgarization of MODAL’s research to sensitize students and businesses: https://modal.lille.inria.fr/xperium/

11. Bibliography

Major publications by the team in recent years

[1] A. Amarioarei, C. Preda. *Approximation for the Distribution of Three-dimensional Discrete Scan Statistic*, in "Methodology and Computing in Applied Probability", September 2013, 14 p. [*DOI :* 10.1007/s11009-013-9382-3], https://hal.inria.fr/hal-01092992

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

[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] 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

[6] M. Giacofci, S. Lambert-Lacroix, G. Marot, F. Picard. *Wavelet-based clustering for mixed-effects functional models in high dimension*, in "Biometrics", March 2013, vol. 69, n^o 1, pp. 31-40 [*DOI :* 10.1111/j.1541-0420.2012.01828.x], http://hal.inria.fr/hal-00782458

[7] J. JACQUES, C. BIERNACKI. *Extension of model-based classification for binary data when training and test populations differ*, in "Journal of Applied Statistics", 2010, vol. 37, n^{o} 5, pp. 749-766, https://hal.archives-ouvertes.fr/hal-00316080

[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] A. LOURME, C. BIERNACKI. *Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins*, in "Computational Statistics", December 2013, vol. 152, n^{o} 3, pp. 371-391, https://hal.inria.fr/hal-00921041

[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

Doctoral Dissertations and Habilitation Theses

[11] C. THÉRY. *Model-based covariable decorrelation in linear regression (CorReg). Application to missing data and to steel industry*, Université Lille 1, July 2015, https://hal.archives-ouvertes.fr/tel-01249789

Articles in International Peer-Reviewed Journals

[12] C. BIERNACKI, J. JACQUES. *Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm*, in "Statistics and Computing", June 2015, https://hal.inria.fr/hal-01052447

[13] C. BOUVEYRON, E. CÔME, J. JACQUES. *The discriminative functional mixture model for a comparative analysis of bike sharing systems*, in "The Annals of Applied Statistics", 2015, forthcoming, https://hal.archives-ouvertes.fr/hal-01024186

[14] S. DABO-NIANG, A. LAKSACI, Z. KAID. *On spatial conditional quantile estimation for a functional regressor*, in "AStA Advances in Statistical Analysis", February 2015, https://hal.inria.fr/hal-01206774

[15] S. DABO-NIANG, G.-M. NKIET, S. BOUKA. *Non-parametric level set estimation for spatial data*, in "Advances and Applications in Statistics", September 2015, vol. 46, n^{o} 2, pp. 119 - 158 [*DOI :* 10.17654/ADASAUG2015_119_158], https://hal.inria.fr/hal-01206787

[16] S. DABO-NIANG, C. TERNYNCK, F. CHEBANA, M. ALI BEN ALAYA, T. OUARDA. *Streamflow hydrograph classification using functional data analysis*, in "Journal of Hydrometeorology", October 2015 [*DOI :* 10.1175/JHM-D-14-0200.1], https://hal.inria.fr/hal-01206807

[17] S. DABO-NIANG, A.-F. YAO, C. TERNYNCK. *A new spatial regression estimator in the multivariate context*, in "Comptes rendus de l'académie des sciences, Mathématiques", April 2015 [*DOI :* 10.1016/J.CRMA.2015.04.004], https://hal.inria.fr/hal-01206781

[18] S. IOVLEFF. *Probabilistic Auto-Associative Models and Semi-Linear PCA*, in "Advances in Data Analysis and Classification", September 2015, vol. 9, n^{o} 3, 20 p. [*DOI :* 10.1007/S11634-014-0185-3], https://hal.archives-ouvertes.fr/hal-00734070

[19] R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, G. Govaert. *Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library*, in "Journal of Statistical Software", 2015, forthcoming, https://hal.archives-ouvertes.fr/hal-00919486

[20] M. Marbac, C. Biernacki, V. Vandewalle. *Model-based clustering for conditionally correlated categorical data*, in "Journal of Classification", 2015, vol. 2, n^{o} 32, pp. 145-175 [*DOI :* 10.1007/s00357], https://hal.inria.fr/hal-00787757

[21] S. Poulain, C. Roumier, A. Venet-Caillault, M. Figeac, C. Herbaux, G. Marot, E. DOYE, E. Bertrand, S. GEFFROY, F. LEPRETRE, O. Nibourel, A. DECAMBRON, E. Boyle, A. Renneville, S. Tricot, A. Daudignon, B. Quesnel, P. Duthilleul, C. Preudhomme, X. Leleu. *Genomic landscape of CXCR4 mutations in Waldenstrom's Macroglobulinemia*, in "Clinical Cancer Research", October 2015, vol. 21, n^{o} 22 [*DOI :* 10.1158/1078-0432.CCR-15-0646], https://hal.inria.fr/hal-01230058

Articles in National Peer-Reviewed Journals

[22] K. Benabdeslem, C. Biernacki, M. Lebbah. *Les trois défis du Big-Data: Eléments de reflexion*, in "Statistique et Société", 2015, vol. 3, n^{o} 1, pp. 19-22, https://hal.archives-ouvertes.fr/hal-01198435

Invited Conferences

[23] C. Biernacki. *Clustering: evolution of methods to meet new challenges*, in "Journée thématique Clustering", Issy Les Moulineaux, France, Orange Labs, October 2015, https://hal.archives-ouvertes.fr/hal-01253394

[24] C. Biernacki. *MixtComp software: Model-based clustering/imputation with mixed data, missing data and uncertain data*, in "MISSDATA 2015", Rennes, France, June 2015, https://hal.archives-ouvertes.fr/hal-01253393

[25] C. Biernacki, T. Deregnaucourt, V. Kubicki. *Model-based clustering with mixed/missing data using the new software MixtComp*, in "CMStatistics 2015 (ERCIM 2015)", London, United Kingdom, December 2015, https://hal.archives-ouvertes.fr/hal-01249829

[26] S. Dabo-Niang. *Spatial risk estimation and application to biomedical data*, in "Tunisian Association of Statistics and Applications Conference", Hammamet, Tunisia, March 2015, https://hal.inria.fr/hal-01206865

[27] S. Dabo-Niang. *Spatial Risk estimation and Applications*, in "African Women in Mathematics Association Conference", Navaisha, Kenya, July 2015, https://hal.inria.fr/hal-01206866

[28] J. Jacques. *Clustering multivariate ranking data*, in "60th World Statistics Congress – ISI2015", Rio de Janeiro, Brazil, July 2015, https://hal.inria.fr/hal-01241256

[29] J. Jacques. *The discriminative functional mixture model for a comparative analysis of bike sharing systems*, in "CMStatistics 2015, 8th International Conference of the ERCIM working group on Computational and Methodological Statistics", London, United Kingdom, December 2015, https://hal.inria.fr/hal-01241254

[30] F. Loingeville, J. Jacques, C. Preda, P. Guarini, O. Molinier. *Modèle Linéaire Généralisé Hiérarchique Gamma-Poisson à 3 facteurs aléatoires - Application au contrôle de qualité*, in "47èmes Journées de Statistique", Lille, France, Société Française de Statistique, June 2015, https://hal.archives-ouvertes.fr/hal-01152840

[31] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Model-based clustering of categorical data by relaxing conditional independence*, in "Classification Society Meeting", Hamilton, Ontario, Canada, Mc Master University, June 2015, https://hal.inria.fr/hal-01238334

[32] C. PREDA. *Regression with categorical functional data*, in "8th International Conference of the ERCIM WG on Computational and Methodological Statistics", Londres, United Kingdom, December 2015, https://hal.inria.fr/hal-01251289

[33] V. VANDEWALLE, C. BIERNACKI. *An efficient SEM algorithm for Gaussian Mixtures with missing data*, in "8th International Conference of the ERCIM WG on Computational and Methodological Statistics", Londres, United Kingdom, December 2015, https://hal.inria.fr/hal-01242588

[34] V. VANDEWALLE, C. COZMA, C. PREDA. *Clustering categorical functional data Application to medical discharge letters*, in "8th International Conference of the ERCIM WG on Computational and Methodological Statistics", Londres, France, December 2015, https://hal.inria.fr/hal-01251284

International Conferences with Proceedings

[35] C. BOUVEYRON, J. JACQUES. *Un algorithme EM pour une version parcimonieuse de l'analyse en composantes principales probabiliste*, in "EGC 2015 - 15ème conférence internationale sur l'extraction et la gestion des connaissances", Luxembourg, Luxembourg, January 2015, https://hal.inria.fr/hal-01241262

[36] M. A. HASNAT, J. VELCIN, S. BONNEVAY, J. JACQUES. *Simultaneous Clustering and Model Selection for Multinomial Distribution: A Comparative Study*, in "Intelligent Data Analysis", Saint Etienne, France, October 2015, https://hal.inria.fr/hal-01203561

Conferences without Proceedings

[37] M. BRUNIN, C. BIERNACKI, A. CELISSE. *Compromis précision-temps de calcul et détection de ruptures*, in "6ème Rencontres des Jeunes Statisticiens", Le Teich, France, August 2015, https://hal.inria.fr/hal-01238276

[38] Q. GRIMONPREZ, A. CELISSE, G. MAROT. *Sélection de groupes de variables corrélées par classification ascendante hiérarchique et group-lasso*, in "Sixièmes rencontres des jeunes statisticiens", Le Teich, France, SFdS, August 2015, https://hal.inria.fr/hal-01238253

[39] Q. GRIMONPREZ, A. CELISSE, G. MAROT. *Sélection de groupes de variables corrélées par classification ascendante hiérarchique et group-lasso*, in "47èmes Journées de Statistique", Lille, France, June 2015, https://hal.inria.fr/hal-01238248

[40] S. IOVLEFF. *Rtkpp: Un package pour faire l'interface entre R et la bibliothèque STK++*, in "Quatrièmes Rencontres R", Grenoble, France, June 2015, https://hal.inria.fr/hal-01253792

Scientific Books (or Scientific Book chapters)

[41] C. BIERNACKI. *Mixture models*, in "Choix de modèles et agrégation", J.-J. DROESBEKE, G. SAPORTA, C. THOMAS-AGNAN (editors), Technip, December 2015, https://hal.inria.fr/hal-01252671

[42] C. BIERNACKI, C. MAUGIS. *High-dimensional clustering*, in "Choix de modèles et agrégation, Sous la direction de J-J. DROESBEKE, G. SAPORTA, C. THOMAS-AGNAN Edition: Technip", December 2015, https://hal.archives-ouvertes.fr/hal-01252673

Other Publications

[43] P. ALQUIER, B. GUEDJ. *Bayesian Non-Negative Matrix Factorization*, January 2016, working paper or preprint, https://hal.inria.fr/hal-01251878

[44] A. CELISSE, T. MARY-HUARD. *New upper bounds on cross-validation for the k-Nearest Neighbor classification rule*, August 2015, working paper or preprint, https://hal.inria.fr/hal-01185092

[45] S. DABO-NIANG, A. AMIRI, M. YAHAYA. *Non-parametric recursive density estimation for space-time data-stream*, September 2015, working paper or preprint, https://hal.inria.fr/hal-01206833

[46] S. DABO-NIANG, A. AMIRI, M. YAHAYA. *Non-parametric recursive density estimation for spatial data*, September 2015, working paper or preprint, https://hal.inria.fr/hal-01206814

[47] S. DABO-NIANG, A. BASSENE, A. DIOP. *Conditional tail index estimation for random fields: fixed design case*, September 2015, working paper or preprint, https://hal.inria.fr/hal-01206824

[48] S. DABO-NIANG, A. BASSENE, A. DIOP. *Kernel estimation of conditional tail index and quantile estimation for random fields*, September 2015, working paper or preprint, https://hal.inria.fr/hal-01206829

[49] S. DABO-NIANG, A. DIOP, A. BASSENE. *A note on conditional tail index estimation for random fields*, September 2015, working paper or preprint, https://hal.inria.fr/hal-01206809

[50] S. DABO-NIANG, B. THIAM. *Nonparametric estimation of a regression function for spatial data with errors*, September 2015, working paper or preprint, https://hal.inria.fr/hal-01206832

[51] C. FRIGUET, F. LETUÉ, V. VANDEWALLE. *Table ronde : "pourquoi et comment enseigner l'analyse de données massives (big data)"*, June 2015, 47èmes Journées de Statistique de la SFdS, https://hal.inria.fr/hal-01250812

[52] B. GUEDJ, S. ROBBIANO. *PAC-Bayesian High Dimensional Bipartite Ranking*, November 2015, working paper or preprint, https://hal.inria.fr/hal-01226472

[53] J. HAMON, G. EVEN, R. DASSONNEVILLE, J. JACQUES, C. DHAENENS. *Use of a novel evolutionary algorithm for genomic selection*, January 2015, working paper or preprint, https://hal.inria.fr/hal-01100660

[54] M. A. HASNAT, J. VELCIN, S. BONNEVAY, J. JACQUES. *Opinion mining from Twitter data using evolutionary multinomial mixture models*, September 2015, working paper or preprint, https://hal.inria.fr/hal-01204613

[55] J. KELLNER, A. CELISSE. *A One-Sample Test for Normality with Kernel Methods*, July 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01175237

[56] L. YENGO, J. JACQUES, C. BIERNACKI, M. CANOUIL. *Variable Clustering in High-Dimensional Linear Regression: The R Package clere*, October 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-00940929