IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Université Lille 2

Activity Report 2014

Project-Team MODAL

MOdel for Data Analysis and Learning

IN COLLABORATION WITH: Laboratoire Paul Painlevé

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team MODAL

Keywords: Statistical Learning, Data Analysis, Classification, Visualization

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

1. Members

Research Scientist
Benjamin Guedj [Inria, Researcher, from Nov 2014]

Faculty Members
Christophe Biernacki [Team leader, Univ. Lille I, Professor, HdR]
Alain Celisse [Univ. Lille I, Associate Professor]
Serge Iovleff [Univ. Lille I, Associate Professor]
Julien Jacques [Univ. Lille I, Associate Professor until Aug 2014, Univ. Lyon II, Professor from Sep 2014, HdR]
Guillemette Marot [Univ. Lille II, Associate Professor]
Cristian Preda [Univ. Lille I, Professor, HdR]
Vincent Vandewalle [Univ. Lille II, Associate Professor]

Engineers
Samuel Blanck [Inria until Oct 2014, Univ. Lille II from Nov 2014]
Vincent Kubicki [Inria]
Komi Nagbe [Inria, from Oct 2014]

PhD Students
Michael Genin [Univ. Lille II, until Sep 2014]
Maxime Brunin [Univ. Lille I, from Oct 2014]
Quentin Grimonprez [Inria]
Jérémie Kellner [Univ. Lille I]
Florence Loingeville [AGLAE]
Matthieu Marbac-Lourdelle [Inria, until Sep 2014]
Clément Théry [ARCELOR-MITTAL]
Loïc Yengo [CNRS, until Sep 2014]

Post-Doctoral Fellow
Alexandru Amarioarei [Univ. Lille I]

Visiting Scientists
Emmanuel Chazard [Univ. Lille II, until Apr 2014]
Sophie Dabo [Univ. Lille III]
Olivier Delrieu [PGXis]

Administrative Assistant
Corinne Jamroz [Inria]

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 (heterogeneous structured data), which are still almost ignored in the literature. Application domains are numerous (credit scoring, marketing,...), but MODAL favors applications related to biology and medicine. Members of the team are already experienced in these directions with complementary skills.

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 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. Application domains

Potential application areas of statistical modeling for heterogeneous data are extensive but some particular areas are identified. For historical reasons and considering the background of the team members, MODAL is mainly focused on biological applications where new challenges in high throughput technologies are opened. In addition, other secondary applications areas are considered in industry, retail, credit scoring and astronomy. Several contacts and collaborations are already established with some partners in these application areas and are described in Sections 7 and 8.

5. New Software and Platforms

5.1. BlockCluster package for co-clustering

Participants: Serge Iovleff, Vincent Kubicki.

BlockCluster is an R package on top of the coclust C++ library. Maintenance of the CRAN package (http://cran.r-project.org/web/packages/blockcluster/index.html) and user assistance on the forum have been ensured.

5.2. clere package for high dimensional regression

Participants: Christophe Biernacki, Loïc Yengo, Julien Jacques.

The clere package for R proposes variable clustering in high dimensional linear regression. It is available on CRAN (http://cran.r-project.org/web/packages/clere/index.html) and now submitted to an international journal dedicated to software [52].

5.3. Clustericat package for correlated categorical variable

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

Clustericat is an R package for model-based clustering of categorical data. In this package, the Conditional Correlated Model (CCM), published in 2014 [24], 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. It is available online on Rforge (https://r-forge.r-project.org/R/?group_id=1803).

5.4. CoModes package for correlated categorical variables

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

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 [49], 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).

5.5. CorReg package for correlated variables in regression

Participants: Christophe Biernacki, Clément Théry.

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. The CorReg package is now available on CRAN (http://cran.r-project.org/web/packages/CorReg/index.html) and graphical functionalities have been added in 2014. It has been presented to a conference [32] and is currently written as a research paper [51]. It is a joint work with Gaétan Loridant.

5.6. HDPenReg package for penalized regressions in high dimension

Participants: Quentin Grimonprez, Serge Iovleff.

HDPenReg is an R-package based on a C++ code dedicated to the estimation of regression model with l1-penalization. It is now available on CRAN (http://cran.r-project.org/web/packages/HDPenReg/index.html). More cross-validation options were added. Maintenance in 2014 concerned bugs correction and documentation updates.

5.7. FunFEM package for functional data

Participant: Julien Jacques.

FunFEM package for R proposes a clustering tool for functional data. The model-based algorithm clusters the functional data into discriminative subspaces. It is available on CRAN (http://cran.r-project.org/web/packages/funFEM/index.html).

5.8. FunHDDC package for functional data

Participant: Julien Jacques.

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. It is available on CRAN (http://cran.r-project.org/web/packages/funHDDC/index.html).

5.9. Galaxy-Modal platform

Participants: Samuel Blanck, Guillemette Marot.

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 infrastruture (http://cloud.france-bioinformatique.fr/). An APP repository with Galaxy-Modal source code has been created (reference : Galaxy - MPAgenomics)

5.10. metaMA package for meta-analysis of microarray data

Participant: Guillemette Marot.

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.

This software is routinely used by biologists from INRA, Jouy en Josas (it has been included in a local analysis pipeline) but its diffusion on the CRAN (http://cran.r-project.org/web/packages/metaMA/index.html) makes it available to a wider community, as attested by the citations of publications related to the methods implemented in the software.

Maintenance in 2014 concerned documentation updates and users assistance.

5.11. metaRNASeq package for meta-analysis of RNA-Seq data

Participant: Guillemette Marot.

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. metaRNASeq is now available on CRAN (http://cran.r-project.org/web/packages/metaRNASeq/index.html).

5.12. 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, submitted for publication in 2014 [48], 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).

5.13. MIXMOD and Rmixmod package for mixed data

Participants: Vincent Kubicki, Christophe Biernacki, Serge Iovleff.

MIXMOD (MIXture MODelling) is an important software for the MODAL team since it concerns its main topics: model-based supervised, unsupervised and semi-supervised classification for various data situations. MIXMOD is now a well-distributed software with over 250 downloads/month are recorded for several years. MIXMOD is written in C++ (more than 10 000 lines) and distributed under GNU General Public License. Several other institutions participate in the MIXMOD development since several years: CNRS, Inria Saclay-Île de France, Université de Franche-Comté, Université Lille 1. The software already benefits from several APP depositions and an R package (Rmixmod) has been associated to MIXMOD in 2012. In 2014, it has led to publication in an international journal dedicated to software [23].

5.14. MixtComp package for full mixed data

Participants: Vincent Kubicki, Christophe Biernacki, Serge Iovleff.

MixtComp (Mixture Computation) is another important software fot the MODAL team since it concerns model-based clustering for mixed data. Main difference with the MIXMOD/Rmixmod software is that MixtComp's architecture is able to integrate easily and quickly all new univariate models, under the conditional independence assumption, that will be sequentially available from researches of the MODAL team or others. Currently, central architecture of MixtComp is built and three models (Gaussian, multinomial, Poisson) are implemented with ability to natively manage missing data (completely or by interval). MixtComp stands both as a C++ library and an R package. The code is currently developed internally, and has been field-tested through two contracted partnerships.

5.15. MPAgenomics package for multi-patient analysis of genomic markers

Participants: Quentin Grimonprez, Guillemette Marot, Alain Celisse.

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.

MPAgenomics is now available on CRAN (http://cran.r-project.org/web/packages/MPAgenomics/index.html). New segmentation methods were added to MPAgenomics. Maintenance in 2014 concerned bugs correction, documentation updates and code factorization.

5.16. RankCluster package to cluster ranking data

Participants: Christophe Biernacki, Quentin Grimonprez, Julien Jacques.

Rankcluster package for R proposes a clustering tool for ranking data. Multivariate and partial rankings can be also taken into account. It is available on CRAN (http://cran.r-project.org/web/packages/Rankcluster/index.html).

Rankcluster now supports tied ranking data. Maintenance in 2014 concerned bugs correction, documentation updates and addition of parallelism.

5.17. rtkpp package: STK++ Integration To R Using Rcpp

Participant: Serge Iovleff.

rtkpp is the integration of the library STK++ (see 5.18) into R. It is using Rcpp. Some functionalities of the Clustering project provided by the library are available in the R environment as R functions. The rtkpp package includes the header files from the STK++ library (currently version 0.8.2). Thus users do not need to install STK++ itself in order to use it. rtkpp is licensed under the GNU GPL version 2 or later and available on CRAN (http://cran.r-project.org/web/packages/rtkpp/index.html).

5.18. STK++ release 0.8.4: The Statistical ToolKit

Participant: Serge Iovleff.

STK++ 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. STK++ is licensed under the GNU LGPL version 2 or later. See: http://www.stkpp.org/

6. New Results

6.1. Highlights of the Year

Thanks to the development technological action MPAGenomics, the team has created one of the first french instances of Galaxy publicly available on the French Bioinformatics cloud. This instance is original as it offers complex statistical tools for genomic data analysis in a user-friendly interface (see 5.9).
The team obtained bilateral contracts with companies as Auchan or RougeGorge thanks to its just emerging, but promising, clustering software MixtComp (see 5.14), dedicated to full mixed and missing data.

6.2. Model for conditionally correlated categorical data

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

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 has been now accepted in an international journal [24]. Furthermore, an R package (Clustericat) is available on CRAN (see 5.3).

6.3. 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. An article has been now submitted to an international journal [49]. Furthermore, an R package (CoModes) is available on Rforge (see 5.4).

6.4. 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. An article has been presented to an international conference [48] and has been also submitted to an international journal [50]. Furthermore, an R package (MixCluster) is available on Rforge (see 5.12).

6.5. Mixture of Gaussians with Missing Data

Participants: Christophe Biernacki, Vincent Vandewalle.

The generative models allow to handle missing data. This can be easily performed by using the EM algorithm, which has a closed form M-step in the Gaussian setting. This can for instance be useful for distance estimation with missing data. It has been proposed to improve the distance estimation by fitting a mixture of Gaussian distributions instead of a considering only one Gaussian component [16]. This is a joint work with Emil Eirola and Amaury Lendrasse .

6.6. Clustering and variable selection in regression

Participants: Christophe Biernacki, Loïc Yengo, Julien Jacques.

A new framework is proposed to address the issue of simultaneous linear regression and clustering of predictors where regression coefficients are assumed to be drawn from a Gaussian mixture distribution. Prediction is thus performed using the conditional distribution of the regression coefficients given the data, while clusters are easily derived from posterior distribution in groups given the data. This work is now published in [27]. Furthermore, an R package (clere) is available on Rforge (see 5.2) and an improved version of the initial model has been submitted to an international journal [52].

6.7. Model-based clustering for multivariate partial ranking data

Participants: Christophe Biernacki, Julien Jacques.

The first model-based clustering algorithm dedicated to multivariate partial ranking data is now published in an internation journal [19]. This is an extension of the (ISR) model for ranking data published in 2013. The proposed algorithm has allowed to exhibit regional alliances between European countries in the Eurovision contest, which are often suspected but never proved.

6.8. Generative models for correlated variables in regression

Participants: Christophe Biernacki, Clément Théry.

Linear regression outcomes (estimates, prevision) are known to be damaged by highly correlated covariates. However most modern datasets are expected to mechanically convey more and more highly correlated covariates due to the global increase of the amount of variables they contain. We propose to explicitly model such correlations by a family of linear regressions between the covariates. It leads to a particular generative model through the distribution explicitly introduced between correlated covariates. It has been presented to a conference [32] and is currently written as a research paper [51]. Furthermore, an R package (CorReg) is available on CRAN (see 5.5). Extension is now available for missing covariables also. It is a joint work with Gaétan Loridant.

6.9. 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 currently in revision in an international journal [38].

6.10. Clustering for functional data into discriminative subspaces

Participant: Julien Jacques.

This is a joint work with Charles Bouveyron (Paris 5) and Etienne Côme (Inrets).

A model-based clustering method for time series has been developed, based on a discriminative functional mixture model which allows the clustering of the data in a functional subspace. This model presents the advantage to be parsimonious and can therefore handle long time series. This model has been used for analyzing different bike sharing systems In Europe.

6.11. Degeneracy in multivariate Gaussian mixtures

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, we derive a simple non-asymptotic stochastic lower bound on variances. We prove also that maximizing the likelihood under this data-driven constraint leads to consistent estimates. This work has been presented to a conference [31]. This is a joint work with Gwënaelle Castellan.

6.12. Auto-Associative Models

Participant: Serge Iovleff.

Auto-Associative models cover a large class of methods used in data analysis, among them are for example the famous PCA and the auto-associative neural networks. We describe the general properties of these models when the projection component is linear and we propose and test an easy to implement Probabilistic Semi-Linear Auto-Associative model in a Gaussian setting. This work is now published in [18].

6.13. Resampling and density estimation

Participant: Alain Celisse.

We characterized the behavior of cross-validation (Lpo) in density estimation with the L^2-loss. We considered two aspects: risk estimation and model selection. For the first one, we settled leave-one-out is optimal. On the contrary for the second one, we provided the first guidelines toward an optimal choice of the parameter p. In particular, this choice depends on the convergence rate of the best estimator in the family we consider.

6.14. Resampling and classification

Participant: Alain Celisse.

This is a joint work with Tristan Mary-Huard (INRA).

We extended known results about leave-one-out to the case of leave-p-out for the k-nearest neighbor estimator in classification with the 0-1 loss. In particular, our strategy relies on the relationship between leave-p-out and U-statistics. We derive upper bounds on the moments on the leave-p-out estimator as well as an exponential concentration inequality.

6.15. Kernel change-point

Participants: Alain Celisse, Guillemette Marot.

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

Based on a previous work, we successfully applied kernel methods to change-point detection in the context of next generation sequencing with multivariate complex data. We also provided greatly improved algorithm in terms of computational complexity (both in time and space). With very huge amounts of data, we also suggest a new strategy based on the idea of approximating the Gram matrix by a low-rank matrix, which leads to a linear time complexity algorithm.

6.16. Normality test in RKHS

Participants: Alain Celisse, Jérémie Kellner.

In the kernel method framework, we use the MMD (maximum mean discrepancy) to derive a new goodness-of-fit test that can be used in the RKHS. When applied to the usual R^d setting, our test does not seem too sensitive to any increase on the dimension d unlike other ongoing approaches. With an infinite dimension RKHS, it exhibits a good power for a prescribed level of type-I error control.

6.17. Differential meta-analysis of RNA-seq data from multiple studies

Participant: Guillemette Marot.

This is a joint work with Andrea Rau and Florence Jaffrézic (INRA, Jouy-en-Josas).

An adaptation of meta-analysis methods initially proposed for microarray studies has been proposed for RNA-seq data. The research paper has been published in [26] and the associated R package metaRNASeq is now available on CRAN (see 5.11).

6.18. Multi-patient analysis of genomic markers

Participants: Quentin Grimonprez, Samuel Blanck, Guillemette Marot, Alain Celisse.

Tests performed during Development Technological Action MPAgenomics have shown on real data that it was also important to suggest automatic and appropriate calibrations for parameters in segmentation methods than to look for common markers able to predict patient's response. In the R package MPAgenomics (see 5.15), we have thus proposed two independent pipelines described in [17]. The choice of a given pipeline depends on the heterogeneity degree of studied genomic profiles.

6.19. Scan statistics for dependent data

Participants: Alexandru Amarioarei, Cristian Preda.

Dependent models of type block factors are introduced for scan statistics as an extension of the models based on the independent and identically distributed assumption. Approximations and errors are derived for one and two dimensions. Matlab software has been developed for this purpose.

7. Bilateral Contracts and Grants with Industry

7.1. Arcelor-Mittal

Participants: Christophe Biernacki, Clément Théry.

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 will finish on 2015.

7.2. PGXIS

Participant: Christophe Biernacki.

PGXIS is a UK pharmacogenomics company aiming to discover virtual drugs. Its business model relies on its star technology, named Taxonomy3, a ground breaking mathematical method. Applied to Big Genetic Data, it delivers novel drug targets that are biologically confirmed. These drug targets will drive its drug discovery programmes. This six months contract aims at developping mathematical tool for accelerating convergence rate of Taxonomy3. From a scientific point of view, it corresponds to define specific importance sampling methods related to the Monte Carlo process involved in Taxonomy3.

7.3. RougeGorge

Participants: Christophe Biernacki, Serge Iovleff, Vincent Vandewalle, Vincent Kubicki, Komi Nagbe.

The RougeGorge company sells lingerie item for women. This three months contract aims at defining a new marketing segmentation for customers and also for items. From a scientific point of view, it corresponds to clustering of mixed data, difficulty being provided but the data volume (millions of customers), by the heterogeneity of data (mixed data) and also by many missing data.

7.4. Auchan

Participants: Christophe Biernacki, Serge Iovleff, Vincent Vandewalle, Vincent Kubicki.

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 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.

7.5. Cap Gemini

Participants: Christophe Biernacki, Vincent Vandewalle.

Cap Gemini S.A. is a French multinational corporation headquartered in Paris, with regional activities. It provides IT services and is one of the world's largest consulting, outsourcing and professional services companies with more than 140,000 employees in over 40 countries. The company aims at developping its Big Data ability in regards to its customer needs. A PhD thesis performing specific research to this activity is planned in 2015. In this aim, a preliminary contract has been established since December 2014. It will allow to prepare precisely the research subject.

7.6. PIXEO

Participant: Christophe Biernacki.

PIXEO is a company allowing online comparisons of insurances. A PhD thesis for optimizing the workflow related to this activity is planned in 2015. In this aim, a preliminary contract has been established since October 2014. It will allow to prepare precisely the research subject. It is a work in collaboration with two members of the Dolphin Inria team (Laetitia Jourdan and Marie-Eléonore Marmion).

7.7. AGLAE

Participants: Julien Jacques, Cristian Preda, Florence Loingeville.

AGLAE aims to improve analyses, especially chemical and microbiological, of water and other matrices of the environment. In the context of the Ph.D. of Florence Loingeville, we work on ANOVA models for counting data.

7.8. Alicante

Participants: Julien Jacques, Cristian Preda, Vincent Vandewalle.

ALICANTE develops applications and tools for data coming from health domain. As a participant of the ClinMine ANR project, ALICANTE and GHICL (Groupe Hospitalier de l'Institut Catholique de Lille) provide us well-structured data for clustering hospital stays.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Collaborations within SIRIC

Participants: Guillemette Marot, Alain Celisse.

SIRIC (Site of integrated research in Cancerology) ONCOLille has been created during "Plan Cancer 2". More information about it can be found at http://www.canceropole-nordouest.org/qui-sommes-nous/le-cancer-en-region/le-siric-oncolille.html. Collaborations established through common articles or funding proposals writings with members of MODAL concern the following teams:

Univ. Lille 2, Functional and structural genomics, M. Figeac

CHRU Lille, Hematology laboratory, C. Preudhomme

CNRS, UMR8161, IBL (Institute of Biology of Lille), O. Pluquet

Inserm UMR837 - Team 5, I. Van Seuningen

8.1.2. Other collaborations

Institut Pasteur Lille, Transcriptomics and Applied Genomics, D. Hot (**Participant:** G. Marot)

Inserm U1011, J. Eeckhoute (**Participants:** G. Marot, A. Celisse)

Registre Regional des Cancers de Lille et sa Region, Dr. Karine Ligier (**Participant:** C. Preda)

8.2. National Initiatives

8.2.1. 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.

8.2.2. Working groups

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

Guillemette Marot belongs to the StatOmique working group http://vim-iip.jouy.inra.fr:8080/statomique/

8.3. International Initiatives

8.3.1. Inria Associate Teams

Associate Team acronym: SIMERGE (Statistics Inference for the Management of Extreme Risks and Global Epidemiology)

Principal investigator (Inria): Stéphane Girard Mistis, Inria Grenoble Rhône-Alpes, France.

Principal investigator (Main team): Abdou Kâ Diongue LERSTAD, Université Gaston Berger, Sénégal.

Other participants: Laboratory EQUIPPE (Economie QUantitative Intégration Politiques Publiques Econométrie), Univ. Lille 1, 2 and 3, MODAL, IRD (Institut de Recherche pour le Développement), Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE), Dakar, Sénégal.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. general chair, scientific chair

Alain Celisse, Julien Jacques and Guillemette Marot organized and were the scientific chairs of the workshop Kernel Methods for Big data, 31 March - 2 April 2014 http://math.univ-lille1.fr/~jacques/KernelMethod-Lille-2014.html

Guillemette Marot was the general chair and the scientific chair of the first session of the scientific day organized by Bilille platform about bioinformatics around integrative biology, Dec 17, 2014 http://www.lifl.fr/~touzet/PPF/integrative.html

Cristian Preda was member of the Scientific Committee of the 7-th international Workshop on Applied Probability (Antalya, 16-19 June, 2014).

9.1.1.2. member of the organizing committee

Since 12, C. Biernacki is the president of the data mining and learning group of the French statistical association (SFdS) http://www.sfds.asso.fr/.

Sophie Dabo-Niang co-organized the session "Functional Statistics and Hydrology" of the international conference "Statistics and Hydrology",in november 2014, Masdar Institute, Abu-Dhabi.

Alain Celisse belongs to the “Mathematical statistics” board of The French Statistical Association (SFDS)

Julien Jacques belongs to the “Statistics and Image” board of The French Statistical Association (SFDS) and animated the probability and statistics seminar of the Laboratory of mathematics “Painlevé” of U. Lille 1.

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

9.1.2. Journal

9.1.2.1. member of the editorial board

Christophe Biernacki is an associate editor, since 10, of the journal Case Studies in Business, Industry and Government Statistics (CSBIGS) http://www.bentley.edu/centers/csbigs.

Cristian Preda is an associate editor (since 2013) of the Journals : Methodology and Computing in Applied Probability (http://www.springer.com/statistics/journal/11009) and Romanian Journal of Mathematics and Computer Science (http://www.rjm-cs.ro)

9.1.2.2. reviewer

Alain Celisse is reviewer for numerous top-level statistical journals: Annals of Statistics, Electronic journal of Statistics, Biometrika,JSPI...

Julien Jacques has reviewed papers for Statistics and Computing, Journal of Applied Statistics, Computational Statistics and Data Analysis, Communications in Statistics, European Journal of Operational Research.

Guillemette Marot has reviewed papers for Statistical Applications in Genetics and Molecular Biology, BMC Bioinformatics and Journal of Bioinformatics Research Studies

Vincent Vandewalle has reviewed papers for Pattern Recognition, Advances in Data Analysis and Classification and International journal of computer mathematics

9.1.3. Invited Talks

Christophe Biernacki gave two invited talks of two hours each to the "Journées d'Etudes en Statistique" in October 2014 at Frejus (http://www.sfds.asso.fr/23-Les_Journees_dtudes_en_Statistique_JES). Subjects were "mixture models" and "high dimensional clustering", both from the general topic model choice and agregation.

Cristian Preda was invited as speaker in the special session on Statistics in functional and Hilbert spaces, ERCIM 2014, 7th International Conference of the ERCIM working group on Computational and Methodological Statistics, Pisa, Italy, December 2014.

The other invited talks are included into the bibliography.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

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

Vincent Vandewalle is head of the DUT Statistique et Informatique Décisionnelle, http://iut.univ-lille2.fr/fr/statistique-et-informatique-decisionnelle.html

Licence: A. Celisse, Mathematics for computer science, 48h, L1, U. Lille 1, France

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, Discrete mathematics, 42h, L1, U. Lille 1, France

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

Licence: C. Preda, Probability, 40h, L1, U. Lille 1, France

Licence: C. Preda, Inferential Statistics, 50h, L1, U. Lille 1, France

Licence : V. Vandewalle, Simulation Techniques, 16h, L1 , U. Lille 2, France

Licence : V. Vandewalle, Descriptive statistics, 62h, L1 , U. Lille 2, France

Licence : V. Vandewalle, Probabilities, 80h, L1 , U. Lille 2, France

Licence: A. Celisse, Mathematics for computer science, 48h, L2, U. Lille 1, France

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

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

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

Licence : V. Vandewalle, Project management, 9h, L2, U. Lille 2, France

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

Licence : J. Jacques, Inferential Statistics, 50h, L3, U. Lille 1,France

Licence : V. Vandewalle, Data analysis, 30h, L3, U. Lille 2, France

Licence : Q. Grimonprez, Probability, 4h, L3, U. Lille 1, France.

Master : C. Biernacki, Mathematical statistics, 60h, M1, U. Lille 1, France

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

Master : Q. Grimonprez, Data Analysis, 36h, M1, U. Lille 1, France

Master : Q. Grimonprez, Classification, 8h, M1, U. Lille 1, France

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

Master: J. Jacques, Data Analysis, 40h, M1, U. Lille 1,France

Master: J. Jacques, Statistical Modelling, 30h, M1, U. Lille 1,France

Master: M. Marbac-Lourdelle, Data Analysis, 48h, M1, U. Lille 1, France

Master: G. Marot, Biostatistics, 49h, M1, U. Lille 2, France

Master: G. Marot, Coaching project, 10h, M1, U. Lille 2, France

Master: C. Preda, Data Analysis, 40h, M1, U. Lille 1, France

Master: C. Thery, Linear Models, 12h, M1, U. Lille 1, France

Master : C. Biernacki, Data analysis, 97.5h, M2, U. Lille 1, France

Master : C. Biernacki, Analysis of variance and experimental design, 22.5h, M2, U. Lille 1, France

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

Master: A. Celisse, Statistical theory, 30h, M2, U. Lille 1, France

Master: B. Guedj, Statistical Learning and Data Mining, 8h, M2, ENSAE ParisTech, France

Master: J. Jacques, Time Series, 25h, M2, U. Lille 1, France.

Master: C. Preda, Biostatistics, 12h, M2, U. Lille 1, France

Master: C. Preda, Functional data analysis, 12h, M2, U. Lille 1, France

Doctorat: G. Marot, Data Analysis with R, 18h, U. Lille 2, France

9.2.2. Supervision

PhD : Loïc Yengo, Contribution to variable clustering in high dimensional linear regression models, Univ. Lille 1, May 28th, 2014, J. Jacques, C. Biernacki

PhD : Alexandru Amarioarei, Approximations for Multidimensional Discrete Scan Statistics, Univ. Lille 1, September 15th, 2014, C. Preda (https://ori-nuxeo.univ-lille1.fr)

PhD : Matthieu Marbac-Lourdelle, Model-based clustering for categorical and mixed data sets, Univ. Lille 1, September 23rd, 2014, C. Biernacki, V. Vandewalle

PhD in progress : Clément Thery, Classification in high dimension, from 2011, C. Biernacki

PhD in progress : Florence Longeville, Analysis of variance with nested factors for counting data - Application to control quality, from Dec 1st, 2012, J. Jacques, C. Preda

PhD in progress : Quentin Grimonprez, Detection of change points and peaks in high dimension, from Oct 1st, 2013, A. Celisse, G. Marot, J. Jacques

PhD in progress : Jérémie Kellner, Generative models and kernel methods Jeremie, from Oct 1st, 2013, A. Celisse, C. Biernacki

PhD in progress : Maxime Brunin, The computation time/accuracy trade-off in statistical learning, from Oct 1st, 2014, A. Celisse, C. Biernacki

9.2.3. Juries

Alain Celisse was a jury member for one associate professor competition.

Sophie Dabo-Niang was a referee and member of the jury of

- Mohamed Badaoui thesis's, March, 2014, Oujda, Maroc
- Tamaro Johng-Ay thesis's, April, 2014, Pau, France
- Abdourahmane Diallo thesis's, December 2014, Marseille, France

Julien Jacques was reviewer of the Ph.D. of Damien McPartland (University College Dublin) and Anastasios Bellas (University Paris 1) and a jury member for one associate professor competition.

Guillemette Marot was a jury member for two associate professor competitions.

9.3. Popularization

- **Fête de la Science**

  Participant: Vincent Kubicki
  Fête de la Science is a series of presentations in high schools. The objective is to meet the students and expose them to the research environment and the applications of the work done by the team.

- **XPérium Lille 1**

  Participants: Quentin Grimonprez, Vincent Kubicki, Samuel Blanck, Maxime Brunin, Christophe Biernacki, Serge Iovleff, Julien Jacques, Vincent Vandewalle
  Vulgarization of MODAL's research to sensitize high school students and members of Lille 1 university: https://modal.lille.inria.fr/xperium/

10. 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$^{\mathrm{o}}$ 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$^{\mathrm{o}}$ 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] A. AMARIOAREI. *Approximations for Multidimensional Discrete Scan Statistics*, Université de Lille 1, September 2014, https://hal.inria.fr/tel-01105214

[12] M. MARBAC-LOURDELLE. *Model-based clustering for categorical and mixed data sets*, université lille 1, September 2014, https://tel.archives-ouvertes.fr/tel-01076418

Articles in International Peer-Reviewed Journals

[13] A. AMARIOAREI, C. PREDA. *Approximations for two-dimensional discrete scan statistics in some block-factor type dependent models*, in "Journal of Statistical Planning and Inference", January 2014, vol. 151-152, 14 p. [*DOI :* 10.1016/J.JSPI.2014.05.002], https://hal.inria.fr/hal-01092993

[14] S. DABO-NIANG, S. ALI OULD ABDI, A. OULD ABDI, A. DIOP. *Consistency of a nonparametric conditional mode estimator for random fields*, in "Statistical Methods and Applications", 2014, 21 p. [*DOI :* 10.1007/S10260-013-0239-2], https://hal.inria.fr/hal-00921178

[15] S. DABO-NIANG, L. HAMDAD, C. TERNYNCK, A.-F. YAO. *A kernel spatial density estimation with applications to spatial clustering and Monsoon Asia Drought Atlas analysis*, in "Stochastic Environmental Research and Risk Assessment", June 2014, pp. 2075-2099, http://hal.univ-lille3.fr/hal-01094743

[16] E. EIROLA, A. LENDASSE, V. VANDEWALLE, C. BIERNACKI. *Mixture of Gaussians for Distance Estimation with Missing Data*, in "Neurocomputing", May 2014, vol. 131, pp. 32-42, https://hal.inria.fr/hal-00921023

[17] Q. GRIMONPREZ, A. CELISSE, S. BLANCK, M. CHEOK, M. FIGEAC, G. MAROT. *MPAgenomics : An R package for multi-patients analysis of genomic markers*, in "BMC Bioinformatics", December 2014, vol. 15, 4 p. [*DOI :* 10.1186/S12859-014-0394-Y], https://hal.inria.fr/hal-00933614

[18] S. IOVLEFF. *Probabilistic Auto-Associative Models and Semi-Linear PCA*, in "Advances in Data Analysis and Classification", September 2014, 20 p. , https://hal.archives-ouvertes.fr/hal-00734070

[19] J. JACQUES, C. BIERNACKI. *Model-based clustering for multivariate partial ranking data*, in "Journal of Statistical Planning and Inference", June 2014, vol. 149, pp. 201-217, https://hal.inria.fr/hal-00743384

[20] J. JACQUES, Q. GRIMONPREZ, C. BIERNACKI. *Rankcluster: An R package for clustering multivariate partial rankings*, in "The R Journal", June 2014, vol. 6, n^{o} 1, 10 p. , https://hal.archives-ouvertes.fr/hal-00840692

[21] J. JACQUES, C. PREDA. *Functional data clustering: a survey*, in "Advances in Data Analysis and Classification", January 2014, vol. 8, n^{o} 3, 24 p. [*DOI :* 10.1007/S11634-013-0158-Y], https://hal.inria.fr/hal-00771030

[22] J. JACQUES, C. PREDA. *Model-based clustering for multivariate functional data*, in "Computational Statistics and Data Analysis", June 2014, vol. 71, pp. 92-106, https://hal.archives-ouvertes.fr/hal-00713334

[23] 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", December 2014, forthcoming, https://hal.archives-ouvertes.fr/hal-00919486

[24] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Model-based clustering for conditionally correlated categorical data*, in "Journal of Classification", 2015, n^{o} 32, 33 p. [*DOI :* 10.1007/S00357], https://hal.inria.fr/hal-00787757

[25] N. MARTIN, C. SALAZAR-CARDOZO, C. VERCAMER, L. OTT, G. MAROT, P. SLIJEPCEVIC, C. ABBADIE, O. PLUQUET. *Identification of a gene signature of a pre-transformation process by senescence evasion in normal human epidermal keratinocytes*, in "Molecular Cancer", 2014, vol. 13, n^{o} 1, 151 p. [*DOI :* 10.1186/1476-4598-13-151], https://hal.inria.fr/hal-01011012

[26] A. RAU, G. MAROT, F. JAFFRÉZIC. *Differential meta-analysis of RNA-seq data from multiple studies*, in "BMC Bioinformatics", 2014, vol. 15, n^{o} 1, 91 p. [*DOI :* 10.1186/1471-2105-15-91], https://hal.inria.fr/hal-00978902

[27] L. YENGO, J. JACQUES, C. BIERNACKI. *Variable clustering in high dimensional linear regression models*, in "Journal de la Société Française de Statistique", 2014, vol. 155, n^{o} 2, 19 p. , https://hal.archives-ouvertes.fr/hal-00764927

Invited Conferences

[28] S. DABO-NIANG. *Statistical modeling of spatial functional data: application to biomedical data*, in "ISNPS conférence", Cadiz, Estonia, June 2014, http://hal.univ-lille3.fr/hal-01094757

[29] S. DABO-NIANG, S. GUILLAS, C. TERNYNCK. *Efficiency in functional nonparametric models withautoregressive errors*, in "ERCIM", Pise, Italy, December 2014, http://hal.univ-lille3.fr/hal-01094762

[30] J. JACQUES, C. BIERNACKI. *Clustering multivariate ordinal data*, in "ERCIM 2014, 7th International Conference of the ERCIM working group on Computational and Methodological Statistics", Pisa, Italy, December 2014, https://hal.inria.fr/hal-01100630

International Conferences with Proceedings

[31] C. BIERNACKI, G. CASTELLAN. *A Data-Driven Bound on Covariance Matrices for Avoiding Degeneracy in Multivariate Gaussian Mixtures*, in "46° Journées de Statistique", Rennes, France, June 2014, https://hal.inria.fr/hal-01099080

[32] C. THÉRY, C. BIERNACKI, G. LORIDANT. *CorReg : Préselection de variables en régression linéaire avec fortes corrélations*, in "46° journées de statistiques", Rennes, France, SFDS, June 2014, https://hal.inria.fr/hal-01092964

[33] L. YENGO, J. JACQUES, C. BIERNACKI. *Variable Clustering in High Dimensional Probit Regression Models*, in "46èmes Journées de Statistique organisée par la Société Française de Statistique", Rennes, France, 2014, https://hal.archives-ouvertes.fr/hal-01100633

National Conferences with Proceedings

[34] Q. GRIMONPREZ, A. CELISSE, G. MAROT. *Analyse multi-patients de données génomiques*, in "46e Journées de Statistique", Rennes, France, SFDS, June 2014, https://hal.archives-ouvertes.fr/hal-01091476

[35] J. KELLNER, A. CELISSE. *High-dimensional test for normality*, in "Journées des Statistiques", Rennes, France, June 2014, https://hal.inria.fr/hal-01091513

[36] F. LOINGEVILLE, J. JACQUES, C. PREDA, P. GUARINI, O. MOLINIER. *Analyse de variance à 2 facteurs imbriqués sur données de comptage - Application au contrôle de qualité*, in "46e Journées de Statistique", Rennes, France, June 2014, 6 p. , https://hal.archives-ouvertes.fr/hal-00986457

Research Reports

[37] M. ATTOUCH, M. SALEM AHMED, S. DABO-NIANG, A. DIOP. *k-nearest neighbors method estimation of regression function for spatial dependent data*, 2014, https://hal.inria.fr/hal-00943647

[38] C. BIERNACKI, J. JACQUES. *Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm*, July 2014, https://hal.inria.fr/hal-01052447

[39] S. BOUKA, S. DABO-NIANG, G. GAYRAUD, G.-M. NKIET. *Minimax testing in a spatial discrete regression scheme*, 2014, https://hal.inria.fr/hal-00943645

[40] S. DABO-NIANG, C. TERNYNCK, A.-F. YAO. *A new spatial regression estimator in the multivariate context*, 2014, https://hal.inria.fr/hal-00943646

[41] J. KELLNER, A. CELISSE. *New goodness-of-fit tes for normality in RKHS*, Inria, 2014, https://hal.inria.fr/hal-00943669

Other Publications

[42] P. BHATIA, S. IOVLEFF, G. GOVAERT. *blockcluster: An R Package for Model Based Co-Clustering*, December 2014, https://hal.inria.fr/hal-01093554

[43] C. BOUVEYRON, E. CÔME, J. JACQUES. *The Discriminative Functional Mixture Model for the Analysis of Bike Sharing Systems*, July 2014, https://hal.archives-ouvertes.fr/hal-01024186

[44] S. DABO-NIANG, A. LAKSACI, Z. KAID. *On spatial conditional quantile estimation for a functional regressor*, July 2014, http://hal.univ-lille3.fr/hal-01094745

[45] Q. GRIMONPREZ, A. CELISSE, G. MAROT. *Analysis of genomic markers: Make it easy with the R package MPAgenomics*, January 2014, SMPGD 2014, https://hal.archives-ouvertes.fr/hal-01091543

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

[47] J. KELLNER, A. CELISSE. *New normality test in high dimension with kernel methods*, April 2014, https://hal.archives-ouvertes.fr/hal-00977839

[48] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Classification de données mixtes par un modèle de mélange de copules gaussiennes*, February 2014, 46e Journées de Statistique (Rennes, du 2 au 6 juin 2014), https://hal.archives-ouvertes.fr/hal-00940613

[49] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Finite mixture model of conditional dependencies modes to cluster categorical data*, February 2014, https://hal.archives-ouvertes.fr/hal-00950112

[50] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Model-based clustering of Gaussian copulas for mixed data*, May 2014, https://hal.archives-ouvertes.fr/hal-00987760

[51] C. THÉRY, C. BIERNACKI, G. LORIDANT. *Model-Based Variable Decorrelation in Linear Regression*, August 2014, https://hal.archives-ouvertes.fr/hal-01099133

[52] L. YENGO, J. JACQUES, C. BIERNACKI, M. CANOUIL. *Variable Clustering in High-Dimensional Linear Regression: The R Package clere*, February 2014, https://hal.archives-ouvertes.fr/hal-00940929