IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Université Lille 2

Activity Report 2013

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

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, HdR]
Guillemette Marot [Univ. Lille II, Associate Professor]
Cristian Preda [Univ. Lille I, Professor, HdR]
Vincent Vandewalle [Univ. Lille II, Associate Professor]

External Collaborators

Sophie Dabo [Univ. Lille III, Professor, from Jan 2013, HdR]
Olivier Delrieu [PGXis,from Jan 2013]
Thomas Verdebout [Univ. Lille III, Associate Professor, from Jan 2013, HdR]

Engineers

Samuel Blanck [Inria, from Oct 2013]
Vincent Kubicki [Inria, from Oct 2013]

PhD Students

Michael Genin [Univ. Lille II, until 2013]
Julie Hamon [Univ. Lille 1, until 2013]
Alexandru Amarioarei [Univ. Lille I]
Quentin Grimonprez [Inria, granted by D.G.A., from 2013]
Jérémie Kellner [Univ. Lille I, granted by Min. de L'Ens. Sup. et de la Rech., from 2013]
Florence Loingeville [with AGLAE team, from 2013]
Matthieu Marbac-Lourdelle [granted by Inria, from 2011]
Clément Thery [granted by CIFRE ARCELOR-MITTAL, from 2011]
Loic Yengo [CNRS]

Administrative Assistant

Sandrine Meilen [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.

2.2. Highlights of the Year

- The team developed an extended version of the Rmixmod package allowing to cluster simultaneous mixed categorical and continuous data (see Section [Rmixmod package for mixed data]).
- The full understanding of cross-validation procedures in density estimation has been tackled with new results in terms of risk estimation and model selection (Section [Resampling procedures]).

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. Domain

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. Software and Platforms

5.1. Rmixmod package for mixed data

Participants: Christophe Biernacki, Serge Iovleff, Parmeet Bhatia.

MIXMOD (MIXture MODelling) is an important software for the mΘdal 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 2013, Parmeet Bhatia, under scientific supervision of Christophe Biernacki and Serge Iovleff, has developped possibility in Rmixmod to cluster simultaneously continuous and categorical data with the restrictive conditional independence assumption. It is an important first step towards the long term purpose of mΘdal to cluster heterogeneous (or mixed) data sets. It is a joint work with Florent Langrognet, Rémi Lebret, Gilles Celeux and Gérard Govaert.

5.2. RankClust package for rank 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. Available on CRAN.

5.3. 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. Available on CRAN.

5.4. Clustericat package for correlated categorical variable

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

`Clustericat` is a R package for model-based clustering of categorical data. In this package, the model `CCM` [41] where the main conditional dependencies between variables are taken into account is implemented. `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).

5.5. CorReg package for correlated variables in regression

Participants: Christophe Biernacki, Clément Théry.

Databases from the steel industry are often large (very long process with many parameters) and have strong correlations between variables. Some variables may be written directly in terms of other via physical models or related by definition. Moreover the process, which is specific to the type of finished product, conditions most of the process parameters and therefore induces strong correlations between variables. The main idea 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. This research is placed in a steel industry context (Arcelor-Mittal Dunkerque).

The associated CorReg package is now available on Rforge. It is a joint work with Gaétan Loridant.

5.6. AAM

Participant: Serge Iovleff.

A console based program written in C++ abd dedicated to the estimation of the Auto-Associative Models.

5.7. BlockCluster

Participants: Serge Iovleff, Parmeet Bathia.

Serge Iovleff, Parmeet Bathia
BlockCluster: An R package on top of the coclust C++ library.

5.8. HDPenReg

Participants: Quentin Grimonprez, Serge Iovleff.

R-package written in collaboration based on a C++ code dedicated to the estimation of regression model with l1-penalization.

5.9. STK++ release 0.5

Participant: Serge Iovleff.

New release including new functionalities for templated expression evaluation (similar to the Eigen library offer) and a new subproject offering tools for Clustering.

5.10. Funclustering package for R

Participants: Cristian Preda, Julien Jacques.

Funclustering package for R proposes a clustering tool for functional data. Multivariate curves can be also taken into account. Available on CRAN.

5.11. metaMA

metaMA is a specialised software for microarrays. It is a 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.

Guillemette Marot is the main contributor and the maintainer of this package.

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 makes it available to a wider community, as attested by the citations of publications related to the methods implemented in the software.

More information is available on the website http://cran.r-project.org/web/packages/metaMA/

5.12. metaRNASeq

Participant: Guillemette Marot.

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. Guillemette Marot and Andrea Rau are the main contributors of this package and Guillemette Marot is the maintainer of this package.

5.13. MPAGenomics

Participants: Quentin Grimonprez, Guillemette Marot, Alain Celisse.

MPAGenomics is a R package for multi-patients analysis of genomics markers. Its main contributor is Quentin Grimonprez. 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. It is available on the Inria forge and should be released on the R-forge in January.

5.14. SMVar

Participant: Guillemette Marot.

SMVar is a specialised software for microarrays. This R package implements the structural model for variances in order to detect differentially expressed genes from gene expression data. It performs gene expression differential analysis, based on a particular variance modelling. Its main competitor is the Bioconductor R package limma but limma assumes a common variance between the two groups to be compared while SMVar relaxes this assumption.

Guillemette Marot is the main contributor and the maintainer of this package.

More information is available on the website http://cran.r-project.org/web/packages/SMVar/index.html

6. New Results

6.1. Resampling procedures

Participant: Alain Celisse.

The new deep understanding of cross-validation procedures in density estimation has been tackled with new results in terms of risk estimation and model selection [7]. This is the first step towards a fully data-driven and optimal choice of cross-validation strategy.

6.2. Kernel change-point

Participants: Alain Celisse, Guillemette Marot, Morgane Pierre-Jean.

On the basis of theoretical arguments, an empirical analysis has been carried out to assess the influence of the choice of the kernel in the kernel change-point strategy described in [2]. This assessment has been done in the biological context of copy number variation and allele B fraction. Several talks have been given in seminars (SSB seminar in Paris,...) and workshops (JSFDS, SMPGD,...)

6.3. Gaussian process in RKHS

Participants: Alain Celisse, Jérémie Kellner.

Since numerous papers make a Gaussian assumption for observations in the reproducing kernel Hilbert space (RKHS), it is important to be able to assess the validity of this crucial assumption. As long as it has been validated, the Gaussian framework can be further used to infer statistical properties of the population at hand (mean, variance,...).

A statistical test has been designed to address such questions at the RKHS level. It is fully computationally efficient and provides really good power in numerous settings. Theoretical properties for the test statistic have been derived as well.

6.4. Model for conditionally correlated categorical data

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

It is a model-based clustering where categorical data are grouped into conditionally independent blocks. The corresponding block distribution is a parsimonious multinomial distribution where the few free parameters correspond to the most likely modality crossings, while the remaining probability mass is uniformly spread over the other modality crossings. The exact computation of the integrated complete-data likelihood allows to perform the model selection, by a Gibbs sampler, reducing the computing time consuming by parameter estimation and avoiding BIC criterion biases pointed out by our experiments.

This model was presented in a conference [13] with scientific committee and in a seminar [17]. An article will be soon submitted. Furthermore, a R package is currently under development.

6.5. Mixture model for mixed kind of data

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

A mixture model of Gaussian copula allows to cluster mixed kind of data. Each component is composed by classical margins while the conditional dependencies between the variables is modeled by a Gaussian copula. The parameter estimation is performed by a Gibbs sampler. This model was presented in a conference [14]. Some technical points will be developed before providing an article.

6.6. 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 [21]. This is a joined work with Emil Eirola and Amaury Lendrasse .

A parallel work is in progress on the mixture degeneracy when considering mixture of Gaussians with missing data. It have been experimentally noticed that the degeneracy in this case is particularly slow. This behaviour is different from the usual setting of degeneracy with mixture of Gaussians which is usually rather fast. A first attempt of the theoretical characterization of this behaviour around a degenerated solution has been presented at a conference [16].

6.7. Transfert learning in model-based clustering

Participant: Christophe Biernacki.

In many situations one needs to cluster several datasets, possibly arising from different populations, instead of a single one, into partitions with identical meaning and described by similar features. Such situations involve commonly two kinds of standard clustering processes. The samples are clustered traditionally either as if all units arose from the same distribution, or on the contrary as if the samples came from distinct and unrelated populations. But a third situation should be considered: As the datasets share statistical units of same nature and as they are described by features of same meaning, there may exist some link between the samples. We propose a linear stochastic link between the samples, what can be justified from some simple but realistic assumptions, both in the Gaussian and in the t mixture model-based clustering context [26]. This is a joint work with Alexandre Lourme.

6.8. Gaussian Models Scale Invariant and Stable by Projection

Participant: Christophe Biernacki.

Gaussian mixture model-based clustering is now a standard tool to determine an hypothetical underlying structure into continuous data. However many usual parsimonious models, despite their appealing geometrical interpretation, suffer from major drawbacks as scale dependence or unsustainability of the constraints by projection. In this work we present a new family of parsimonious Gaussian models based on a variance-correlation decomposition of the covariance matrices. These new models are stable by projection into the canonical planes and, so, faithfully representable in low dimension. They are also stable by modification of the measurement units of the data and such a modification does not change the model selection based on likelihood criteria. We highlight all these stability properties by a specific geometrical representation of each model. A detailed GEM algorithm is also provided for every model inference. Then, on biological and geological data, we compare our stable models to standard geometrical ones.

This joint work with Alexandre Lourme is now published in [6].

6.9. 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 [28]

6.10. An AIC-like criterion for semi-supervised classification

Participants: Christophe Biernacki, Vincent Vandewalle.

In semi-supervised classification, generative models take naturally into account unlabeled data and parameter estimation can be easily performed through the EM algorithm. However, traditional model selection criteria either does not take into consideration the predictive purpose (AIC or BIC criteria) or involve a high computational cost because of the EM mechanism (cross validation criteria). Alternatively, we propose the penalized model selection criterion AICcond which aims to estimate the predictive power of a generative model by approximating its predictive deviance. AICcond has similar computational cost to AIC, owns good consistency theoretical properties and highlights encouraging behaviour for variable and model selection in comparison to other standard criteria.

This joint work with Gilles Celeux and Gérard Govaert is now published in[16].

6.11. Consistency of a nonparametric conditional mode estimator for random fields

Participant: Sophie Dabo-Niang.

Sophie Dabo-Niang settled the consistency of a nonparametric conditional mode estimator for random fields, Statistical Methods and Applications [9].

6.12. Spatial linear models

Spatial linear models only capture global linear relationships between locations. However, in many circumstances the spatial dependency is not linear. It is, for example, the classical case where one deals with the spatial pattern of extreme events such as in the economic analysis of poverty, in the environmental science,... This leads naturally to consider nonparametric modeling.

6.13. Auto-associative models

Serge Iovleff gave a complete treatment of the Auto-Associative models in the semi-linear case and wrote a software for estimating these models (hal-00734070, version 1).

6.14. BlockCluster

Serge Iovleff has submitted a paper on the BlockCluster package in collaboration with Parmeet Bathia.

6.15. Rmixmod

Serge Iovleff has contributed to a paper submitted to JSS (hal-00919486, version 1) in collaboration with R. Lebret, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert.

6.16. Clustering for functional data

Participants: Julien Jacques, Cristian Preda.

In Jacques & Preda 2014 (CSDA), we propose a model-based clustering algorithm for multivariate functional data, based on multivariate functional principal components analysis. A review on clustering for functional data has also be published in Jacques & Preda 2014 (ADAC). Variable selection in high-dimensional regression Participants: Julie Hamon, Julien Jacques, Clarisse Dhaenens. In the context of genomic analysis, dealing with high-throughput genotyping data, we develop a genetic algorithm which looks for the best subset of variables (of given size) to predict some quantitative feature.

6.17. Wavelet based clustering using mixed effects functional models

Participant: Guillemette Marot.

The paper related to the wavelet based clustering procedure presented in the activity report from MODAL team in 2012 was published in Biometrics [22].

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

Participant: Guillemette Marot.

An adaptation of meta-analysis methods intially proposed for microarray studies has been proposed for RNA-seq data. The R package metaRNASeq is available on the R Forge and the preprint of the paper is available on Arxiv [48].

6.19. Toxoplasma transcription factor TgAP2XI-5 regulates the expression of genes involved in parasite virulence and host invasion

Participant: Guillemette Marot.

The use of peak detection methods implemented in the Bioconductor package `Ringo` has enabled to better understand part of the gene regulation process in T. Gondii parasite. The new findings in Biology have been published in *Walker (2013)*.

7. Bilateral Contracts and Grants with Industry

7.1. 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.

7.2. Banque Accord

Christophe Biernacki gave a one-day course on the Rmixmod and BlockCluster packages to statistical members of the Banque Accord company.

7.3. Hi Duty Free

Participants: Christophe Biernacki, Serge Iovleff.

HiDutyFree had to solve a combinatorial optimization problem for optimizing its costumer service. For this contract we supervise two internships, giving a mathematical treatment of the problem of HiDutyFree and furnish a beta program based on ruby and java for solving it.

7.4. 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.

7.5. Alicante

Participants: Julien Jacques, Cristian Preda, Florence Loingeville.

Alicante is member of the ANR TecSan ClinMine) we obtained for 2014-2018 to work on the path of patients at the hospital.

8. Partnerships and Cooperations

8.1. Regional Initiatives

- Christophe Biernacki: Industrial studies, Arcelor-Mittal (C. Théry)
- Sophie Dabo-Niang:
  - Festival NEXT avec la ROSE DES VENTS : programme Cartes et Cartel du spectacle vivant – stratégies et fréquentation du festival NEXT en Nord Pas de Calais et Belgique (Tournai).
  - SIRIC (Site de Recherche Intégrée en Cancérologie) ONCOLILLE
- Guillemette Marot:
  - Institut Pasteur Lille, Équipe Etudes Transcriptomiques et Génomiques Appliquées, D. Hot
  - Institut Pasteur Lille, Équipe Peste et Yersinia pestis, F. Sebbane
  - Institut de Biologie de Lille, Unité d'approches fonctionnelle et structurale des cancers, O. Pluquet
  - Université Lille 2, Plate-forme de génomique fonctionnelle et Structurale, M. Figeac
  - CHRU Lille, Centre de Biologie Pathologie, Laboratoire d'Hématologie, C. Preudhomme

8.2. National events

- Alain Celisse belongs to the Statistics for Systems Biology group (SSB) in Paris.
- Julien JACQUES organized the first French Summer School in Astrostatitics (Annecy, October 2013).
- Christophe Biernacki co-organized with Gilles Celeux, Gérard Govaert and Florent Langrognet the 4th one-day meeting on Mixmod on September 2013 ($\sim$ 50 participants).
- Guillemette Marot belongs to the StatOmique working group http://vim-iip.jouy.inra.fr:8080/statomique/

8.3. International Research Visitors

8.3.1. Visits of International Scientists

Mahlet Tadesse (University of Georgetown), Mohamed Ben Alaya (INRS, Québec), Aliou Diop (University of Gaston Berger, Senegal), Papa Ngom (University UCAD, Senegal).

8.3.1.1. Internships

Every year the Modal team welcomes numerous internships from various areas: Master 2 (Applied mathematics in Lille 1, Besançon,...), École centrale Lille, École PolytechLille, IUT A,...Some of them are awarded by a grant and then become PhD students (Jérémie Kellner, Quentin Grimonprez, Julie Hamon, Mathieu Marbac-Lourdelle,...).

8.3.2. Visits to International Teams

Julien Jacques was invited to the Working-Group on Model-Based Clustering of Adrian Raftery (Univ. Washington).

9. Dissemination

9.1. Scientific Animation

- Alain Celisse is reviewer for numerous top-level statistical journal: Annals of Statistics, Electronic journal of Statistics, Biometrika,JSPI...He also reports on various French funding proposals (ANR, PEPII, PEPS,...). Alain Celisse is a member of The French Statistical Association (SFDS) and more precisely belongs to the “Mathematical statistics” board.
- C. Biernacki belongs to the program comity of "Extraction et gestion des connaissances" in 2013 and to the program comity of "Journées Françaises de Statistique" in 2013. Since ’10, he is an Associate Editor of the journal “Case Studies in Business, Industry and Government Statistics” (CSBIGS) http://www.bentley.edu/centers/csbigs.

  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/. Since ’11, he is leader of the team “Probability & Statistics” of the Laboratory of mathematics of U. Lille 1 http://math.univ-lille1.fr/. Since ’13 (September), he is co-leader of the Laboratory of mathematics “Painlevé” of U. Lille 1 http://math.univ-lille1.fr/.
- Sophie Dabo-Niang animates a research group on « Spatial Statistics and Archeology » with statisticians and archeologist.
- Guillemette Marot is a member of the organizing committee of seminars from Bilille platform. More information about seminars of the year is available on https://wikis.univ-lille1.fr/bilille/animation.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

- Christophe Biernacki (head of the M2 Ingénierie Statistique et Numérique http://mathematiques.univ-lille1.fr/Formation/):
  - Master 1st year: Mathematical statistics, 60h, coaching project, 10h, M1, U. Lille 1, France
  - Master 2nd year: Data analysis, 97.5h, Analysis of variance and experimental design, 22.5h, coaching internship, 20h, M2, U. Lille 1, France
- Alain Celisse (6 month delegation at Inria):
  - Licence 1st and 2nd year: (96h) Computer Science departement in IUT A, Univ. Lille 1
  - Master 2nd year: Statistical theory (30h) to Applied Mathematics students.
- Sophie Dabo-Niang:
  - Master MIASHS (Mathématiques, Informatique appliquées aux SHS),
  - Master of Statistics of university Gaston Berger (Senegal).
- Serge Iovleff (6 month delegation at Inria):

  - Markov chain theory, Algebra, graphs, languages and automate theory (120h).
- Julien Jacques:
  - Licence 3rd year: Statistique Inférentielle, 50h, École Polytechnique Universitaire de Lille, U. Lille 1,France
  - Master 1st year: Statistique Exploratoire, 40h, École Polytechnique Universitaire de Lille, U. Lille 1,France
  - Master 1st year: Modélisation Statistique, 30h, École Polytechnique Universitaire de Lille, U. Lille 1,France
  - Master 2nd year: Séries Temporelles, 25h, École Polytechnique Universitaire de Lille, U. Lille 1, France.
- Mathieu Marbac-Lourdelle:
  - Licence: Probabilités, 30h, École Polytechnique Universitaire de Lille, U. Lille 1, France.
  - Licence: Statistiques, 34h, École Polytechnique Universitaire de Lille, U. Lille 1, France.
- Guillemette Marot:
  - Licence: Biostatistics, 12h,PACES (équivalent L1), U. Lille 2, France
  - Master : Biostatistics, 45h, M1, U. Lille 2, France
  - Formation permanente: Data Analysis with R, 12h, U. Lille 2, France
  - Master: Internships, 15h, M1, U. Lille 2, France.

9.2.2. Supervision

- Alain Celisse is co-supervising the Ph.D. theses of Jérémie Kellner and Quentin Grimonprez respectively with Christophe Biernacki and with Julien Jacques and Guillemette Marot.
- Christophe Biernacki supervises Clément Thery and co-supervises Loic Yengo, Matthieu Marbac-Lourdelle and Jérémie Kellner.
- Sophie Dabo-Niang supervises:
  - Aladji BASSENE, until 2011. Co-tutelle with Aliou Diop (University Gaston Berger, Senegal)
  - Stéphane BOUKA, until 2012 : Co-tutelle with Guy Martial Nkiet, University of FranceVille, Gabon)
  - Emad Aldeen DRWESH, until 2012 : co-direction with Jerôme Foncel (Lille 3)
  - Camille TERNYNCK, until 2011 :co-direction with Anne-Françoise Yao et Fateh Chebana
  - Mohamed YAHAYA, until 2012 : co-direction with Aboubacar Amiri (Lille 3)
  - Mohamed Ould Yehdhih, until 2013 : Co-tutelle with Aliou Diop (University Gaston Berger,
  - Senegal) and Mohamed Attouch (University Sidi Bel Abbes, Algeria).
- Serge Iovleff has supervised a software development engineer (Parmeet Bathia).
- Julien Jacques is supervizing Julie Hamon who passed her Ph.D. in November, 26th, 2013. He also co-advised Quentin Grimonprez (co-supervision with Guillemette Marot and Alain Celisse) and Florence Loingeville (co-supervision with Cristian Preda) from 2013.
- Guillemette Marot co-supervised (with Alain Celisse) a one-year engineer Morgane Pierre-Jean, who worked on change point detection with kernel methods for genomic data. She also supervised Quentin Grimonprez (ADT MPAGenomics until october 2013 then PhD) and Samuel Blanck (ADT MPAGenomics from 2013).

9.2.3. Juries

- Alain Celisse was a jury member (examinator) at the Ph.D. defense of Van Hahn NGUYEN (Paris-Sud 11) and during the CR2 INRA competition.
- Christophe Biernacki participated to 7 PhD juries in 2013 (1 as an opponent, 5 as a reviewer, 1 as an examinator).
- Sophie Dabo-Niang was in the PhD jury of Karima Kimouche (University of Constantine; June 2013), Ibrahim Sidi Zakari (University of Marrakesh, June 2013) Van Ly Tran (University of Orleans, December, 12, 2013) Aubin N'dri Yao (University of Abidjan, July, 2013).
- Guillemette Marot was a jury member for the CR2 Inria 2013 competition.

9.3. Popularization

Alain Celisse has given a talk in "30 minutes de science" that is proposed to all Inria team members to illustrate the type research carried out within the different teams in Lille. This talk was about kernel change-point detection.

10. Bibliography

Major publications by the team in recent years

[1] S. Arlot, A. Celisse. *Segmentation of the mean of heteroscedastic data via cross-validation*, in "Statistics and Computing", 2010, pp. 1–20, http://www.springerlink.com/content/jq202v115512u26p/

[2] S. Arlot, A. Celisse, Z. Harchaoui. , *Kernel change-point detection*, 2012

[3] C. Biernacki. *Pourquoi les modèles de mélange pour la classification ?*, in "La Revue de Modulad", 2009, vol. 40, pp. 1–22

[4] C. Biernacki, G. Celeux, G. Govaert. *Exact and Monte Carlo Calculations of Integrated Likelihoods for the Latent Class Model*, in "Journal of Statistical and Planning Inference", 2010, n^o 1, pp. 2991–3002

[5] C. Biernacki, J. Jacques. *A generative model for rank data based on sorting algorithm*, in "Computational Statistics and Data Analysis", 2013, n^o 58, pp. 162–176

[6] A. Biernacki. *Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection*, in "Statistics and Computing", in press

[7] A. Celisse. , *Optimal cross-validation in density estimation*, ArXiv, 2013, n^o arXiv:0811.0802v3

[8] 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

[9] S. Dabo-Niang, S. A. Ould-Abdi, A. Diop. *Consistency of a nonparametric conditional mode estimator for random fields*, in "Statistical Methods and Applications", 2013 [*DOI :* 10.1007/s10260-013-0239-2]

[10] M. Giacofci, S. Lambert-Lacroix, G. Marot, F. Picard. *Wavelet-based clustering for mixed-effects functional models in high dimension*, in "Biometrics", 2012, to appear

[11] M. GUEDJ, A. CELISSE, G. NUEL. *kerfdr: A semi-parametric kernel-based approach to local FDR estimations*, in "BMC Bioinformatics", 2009, vol. 84, n^{o} 10, (electronic)

[12] 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

[13] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Modèle de classification de données qualitatives par modes de dépendance conditionnelle*, in "45e Journées de Statistique de la SFDS, Toulouse", 2013

[14] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Modèle de mélange de copules Gaussiennes pour la classification des données hétérogènes*, in "Cinquièmes Rencontres des Jeunes Statisticien-ne-s, Aussois", 2013

[15] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. , *Model-based clustering for conditionally correlated categorical data*, Inria, 2013, n^{o} RR-8232

[16] 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", in press

[17] V. VANDEWALLE, C. BIERNACKI, M. MARBAC. *Modèle de classification de données qualitatives par modes de dépendance conditionnelle*, in "Seminar of probability and statistics, Paris V", 2013

Publications of the year

Articles in International Peer-Reviewed Journals

[18] C. BIERNACKI, A. LOURME. *Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection*, in "Statistics and Computing", December 2013, In press, http://hal.inria.fr/hal-00688250

[19] 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", 2013 [*DOI :* 10.1007/S10260-013-0239-2], http://hal.inria.fr/hal-00921178

[20] S. DABO-NIANG, A.-F. YAO. *Spatial kernel density estimation for functional random variables*, in "Metrika", 2013, vol. 1, pp. 19-52, http://hal.inria.fr/hal-00943638

[21] E. EIROLA, A. LENDASSE, V. VANDEWALLE, C. BIERNACKI. *Mixture of Gaussians for Distance Estimation with Missing Data*, in "Neurocomputing", December 2013, vol. In press, http://hal.inria.fr/hal-00921023

[22] 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

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

[24] J. JACQUES, C. PREDA. *Functional data clustering: a survey*, in "Advances in Data Analysis and Classification", January 2013, 25 p. [*DOI :* 10.1007/S11634-013-0158-Y], http://hal.inria.fr/hal-00771030

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

[26] 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, http://hal.inria.fr/hal-00921041

[27] R. WALKER, M. GISSOT, L. HUOT, T. DILEZITOKO ALAYI, D. HOT, G. MAROT, C. SCHAEFFER-REISS, A. VAN DORSSELAER, K. KIM, S. TOMAVO. *Toxoplasma transcription factor TgAP2XI-5 regulates the expression of genes involved in parasite virulence and host invasion*, in "Journal of Biological Chemistry", 2013, http://hal.inria.fr/hal-00921150

[28] L. YENGO, J. JACQUES, C. BIERNACKI. *Variable clustering in high dimensional linear regression models*, in "Journal de la Société Française de Statistique", 2014, in press, http://hal.inria.fr/hal-00764927

Invited Conferences

[29] S. DABO-NIANG. *Spatial Data Analysis*, in "8th International conference of Sousse ISG", Sousse, Tunisia, 2013, http://hal.inria.fr/hal-00943641

[30] J. JACQUES. *Clustering multivariate ordinal data*, in "20th Summer Working Group on Model-Based Clustering of the Department of Statistics of the University of Washington", Bologna, Italy, July 2013, http://hal.inria.fr/hal-00943733

[31] J. JACQUES. *Model-based clustering for multivariate functional data*, in "ERCIM 2013, 6th International Conference of the ERCIM working group on Computational and Methodological Statistics", London, United Kingdom, December 2013, http://hal.inria.fr/hal-00943732

International Conferences with Proceedings

[32] S. DABO-NIANG. *Exploring spatial non-parametric estimations*, in "Marrakesh International Conference on Probability and Statistics", Marrakech, Morocco, 2013, http://hal.inria.fr/hal-00943642

Conferences without Proceedings

[33] Q. GRIMONPREZ, J. JACQUES, C. BIERNACKI. *Rankclust: An R package for clustering multivariate partial rankings*, in "Deuxième Rencontres R", France, 2013, http://hal.inria.fr/hal-00944005

[34] J. HAMON, C. DHAENENS, G. EVEN, J. JACQUES. *Feature selection in high dimensional regression problems for genomic*, in "Tenth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics", Nice, France, June 2013, http://hal.inria.fr/hal-00839705

[35] J. HAMON, C. DHAENENS, G. EVEN, J. JACQUES. *Modèles mixtes en génétique animale : sélection de variables par optimisation combinatoire*, in "45ème Journées De Statistiques", Toulouse, France, May 2013, http://hal.inria.fr/hal-00839707

Research Reports

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

[37] S. Bouka, S. Dabo-Niang, G. Gayraud, G.-M. Nkiet. , *Minimax testing in a spatial discrete regression scheme*, 2014, http://hal.inria.fr/hal-00943645

[38] 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*, 2013, http://hal.inria.fr/hal-00943643

[39] S. Dabo-Niang, C. Ternynck, A.-F. Yao. , *A new spatial regression estimator in the multivariate context*, 2014, http://hal.inria.fr/hal-00943646

[40] J. Kellner, A. Celisse. , *New goodness-of-fit tes for normality in RKHS*, 2014, http://hal.inria.fr/hal-00943669

[41] M. Marbac, C. Biernacki, V. Vandewalle. , *Model-based clustering for conditionally correlated categorical data*, Inria, February 2013, n^{o} RR-8232, 33 p. , http://hal.inria.fr/hal-00787757

[42] C. Ternynck, M. Ali Ben Alaya, F. Chebana, S. Dabo-Niang, T. Ouarda. , *Flood hydrograph classification using functional data analysis*, 2013, http://hal.inria.fr/hal-00943644

Patents and standards

[43] M. Pierre-Jean, G. Marot, R. Guillem, A. Celisse. , *Change-point detection with kernel methods : application to DNA copy number signals*, 2013, http://hal.inria.fr/hal-00943413

[44] M. Pierre-Jean, G. Marot, G. Rigaill, A. Celisse. , *Détection de ruptures à partir de méthodes à noyaux*, 2013, http://hal.inria.fr/hal-00943423

Other Publications

[45] Q. Grimonprez, A. Celisse, M. Cheok, M. Figeac, G. Marot. , *MPAgenomics : An R package for multi-patients analysis of genomic markers*, 2014, http://hal.inria.fr/hal-00933614

[46] J. Jacques, Q. Grimonprez, C. Biernacki. , *Rankcluster: An R package for clustering multivariate partial rankings*, 2013, http://hal.inria.fr/hal-00840692

[47] M. Marbac, C. Biernacki, V. Vandewalle. , *Classification de données mixtes par un modèle de mélange de copules gaussiennes*, 2014, 46e Journées de Statistique (Rennes, du 2 au 6 juin 2014), http://hal.inria.fr/hal-00940613

[48] A. Rau, G. Marot, F. Jaffrézic. , *Differential meta-analysis of RNA-seq data from multiple studies*, June 2013, http://hal.inria.fr/hal-00834369

[49] L. Rémi, I. Serge, L. Florent, C. Biernacki, G. Celeux, G. Govaert. , *Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library*, 2013, http://hal.inria.fr/hal-00919486

[50] L. Yengo, J. Jacques, C. Biernacki, M. Canouil. , *Variable Clustering in High-Dimensional Linear Regression: The R Package clere*, 2013, http://hal.inria.fr/hal-00940929