IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Université Lille 2

Activity Report 2012

Project-Team MODAL

MOdel for Data Analysis and Learning

IN COLLABORATION WITH: Laboratoire Paul Painlevé

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, Learning and Statistical Methods

Table of contents

Project-Team MODAL

Keywords: Statistical Learning, Data Analysis, Classification, Visualization

Creation of the Project-Team: January 01, 2012 .

1. Members

Faculty Members

Christophe Biernacki [Team leader, Professor at U. Lille 1, UMR 8524, HdR]
Cristian Preda [Professor at U. Lille 1, UMR 8524, HdR]
Alain Célisse [Associate Professor at U. Lille 1, UMR 8524]
Serge Iovleff [Associate Professor at U. Lille 1, UMR 8524]
Julien Jacques [Associate Professor at U. Lille 1, UMR 8524]
Guillemette Marot [Associate Professor at U. Lille 2, EA 2694, chaire Inria]
Vincent Vandewalle [Associate Professor at U. Lille 2, EA 2694]

Engineers

Parmeet Bhatia [ADT grant]
Perrine Boulenger [ADT grant]
Quentin Grimonprez [ADT grant]

PhD Students

Alexandru Amarioarei [MESR grant]
Michael Genin [MESR grant]
Julie Hamon [CIFRE grant]
Matthieu Marbac-Lourdelle [DGA-Inria grant]
Clément Thery [CIFRE grant]
Loic Yengo [Institut Biologique de Lille]

Administrative Assistant

Sandrine Meilen

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 favours applications related to biology and medicine (see Section 4.1). 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 (see Section 3.1) and data visualization through such models (see Section 3.2). 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 ermerging topics, and publicly available specific softwares relying on the proposed new methodologies.

2.2. Highlights of the Year

- The team finished the development of the blockcluster R package, allowing to process efficient and parsimonious generative models on huge data sets for different kinds of variables (see Section 5.2).
- The team developed also a R package of MIXMOD and started to develop a new version for simultaneous mixed categorical and continuous data (see Section 5.1).

3. Scientific Foundations

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) space, 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, 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 modelling 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 chalanges 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 8.1.

5. Software

5.1. Two advances for the MIXMOD software

Participants: Christophe Biernacki, Serge Iovleff, Remi Lebret, Parmeet Bhatia.

MIXMOD (MIXture MODelling) is an important software for the mΘdal team since it concerns its main topics: model-based supervised, unsupervised and semisupervised 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 deposits.

An interface between MIXMOD and R (Rmixmod) has been developed by Rémi Lebret and Serge Iovleff and is now available on the CRAN (http://cran.r-project.org/web/packages/Rmixmod/index.html). We expect now a wide impact of MIXMOD on the growing community familiar with R. A paper related to Rmixmod is submitted to an international journal [34].

Until December 2012, Parmeet Bhatia, under scientific supervision of Christophe Biernacki, is developping possibility in MIXMOD 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.

5.2. The blockcluster package

Participants: Christophe Biernacki, Serge Iovleff, Parmeet Bhatia.

blockcluster is a R package for model-based simultaneous clustering of rows and columns, thanks to an Inria ADT grant (Parmeet Bhatia). It is also developed in collaboration with University of Technology of Compiègne. It offers the ability to structure very large data tables both in lines and columns for different data types (continuous, binary and contingency data). In particular, it opens wide potential applications in biology, marketing, etc. It is available online on CRAN (http://cran.r-project.org/web/packages/blockcluster/index.html) for all major platforms (Linux, MacOS, Windows). It also comes with utility functions to visualize data. A paper related to blockcluster is submitted to an international journal [40].

5.3. Cuvclust package

Participant: Guillemette Marot.

cuvclust is a R package dedicated to model-based curve clustering. Considered models include Functional Clustering Mixed Models (FCMM, ie functional clustering with the presence of functional random effects), but also traditional functional clustering model (FCM, without functional random effects), and functional mixed models (FMM, functional random effects without clustering). Estimation is done by maximum likelihood using the EM algorithm, and two criteria are proposed to select the number of clusters, based on integrated likelihoods.
Guillemette Marot was the main contributor of the beta version of the package during her post-doc. Due to several changes in conception and due to planning of extensions in the package by the other contributors of the package, she decided to become a regular contributor and left the maintenance to Franck Picard.

5.4. MetaMa

Participant: Guillemette Marot.

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 these packages and spent around one year full time for this package between the conception, the implementation, and the documentation. Her PhD advisors (Florence Jaffrézic, Claus-Dieter Mayer, Jean-Louis Foulley) helped her with the conception but she implemented alone the code.

First versions were posted to the CRAN, the official website of the R software, in 2009. New versions for this package were released in August 2011 in order to take into account remarks from the main users (biologists or biostatisticians analysing gene expression data). 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/

5.5. 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.

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

5.6. Tax3 Software

Participants: Serge Iovleff, Remi Lebret.

Tax3 implements a statistical method providing an analytical framework for high dimensional datasets and complex problems combining several variable types: genetics, genomics, biomarkers and phenotypes

5.7. aam Program

Participant: Serge Iovleff.

aam is a console based program dedicated to the estimation of the semi-linear auto-associative models in a gaussian setting. It is written in C++ and used the STK++ library as support.

5.8. STK++

Participant: Serge Iovleff.

STK++ is a multi-platform toolkit written in C++ for creating fast and easy to use data mining programs. It offers a large set of templated class in C++ which are suitable for projects ranging from small one-off projects to complete statistical application suites. A C equivalent would be gsl. However, STK++ is developed in C++ in order to get speed and reusability.

As the aim of STK++ is to aid developers to new developments, it proposes essentially interfaces classes and various concrete helping classes, like arrays, numerical methods (QR, SVD), input and output (csv files), random number generators, etc.

The software is regularly developed for 10 years by Serge Iovleff and it is a work in progress. The version 0.3 has been released. More information is available on the website http://www.stkpp.org/ and source repository is here: https://sourcesup.cru.fr/projects/stk/

5.9. Scan3D

Participants: Alexandru Amarioarei, Cristian Preda.

Scan3D is a C++ sofware for estimating the distribution of the three-dimensional scan statistics for Bernoulli and Poisson models. It implements the most recent approximation methods available, in particular that developed by the authors providing bounds for the approximation errors [39].

6. New Results

6.1. 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, 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 model was used in September for software components data set of Philippe Merle (ADAM Team Inria Lille).

A conference paper [26] and a poster workshop [35] have been presented. A preprint has been also written [45]. Furthermore, an R package is currently under development.

6.2. 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 has been developed in [43]. This is an extension of the (ISR) model for ranking data published in [4]. The proposed algorithm has allowed to exhibit regional alliances between European countries in the Eurovision contest, which are often suspected but never proved.

6.3. A new probability distribution for ordinal data

Participants: Christophe Biernacki, Julien Jacques.

In [21], a probability distribution for univariate ordinal data is proposed from a stochastic dichotomic search algorithm in a sorting table. Interest of this approach is to give a specific model for ordinal data, without any reference to numerical or nominal data, as it is often the case. The resulting distribution is governed by a position and a dispersion parameter, and is easily estimated by using an EM algorithm.

6.4. Clustering and variable selection in regression

Participants: Christophe Biernacki, Julien Jacques, Loic Yengo.

The works presented in [28] address the issue of simultaneous linear regression and clustering of predictors. A new framework is proposed that both sidesteps optimization challenges and improves prediction performance. In that framework, 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.

6.5. Mixture of Gaussians with Missing Data

Participants: Christophe Biernacki, Vincent Vandewalle.

The generative models allow to handle with 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 in [18] to improve the distance estimation by fitting a mixture of Gaussian distribution instead of a considering only one Gaussian component. An extension of the previous work including the high setting has been submitted in Neurocomputing journal. 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. We are working on the theoretical characterization of this behaviour around a degenerated solution.

6.6. Transfer 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 [37]. This is a joint work with Alexandre Lourme.

A book chapter about transfer learning (including clustering, classification and regression) has been also published [37]. It is a joint work with Farid Beninel, Charles Bouveyron, Julien Jacques and Alexandre Lourme.

6.7. 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 work is was presented as a poster to workshop [31] and is also a preprint [41] currently in revision in an international journal. This is a joint work with Alexandre Lourme.

6.8. Decorrelating variables in high dimension for linear regression

Participants: Christophe Biernacki, Clément Thery.

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-regressions 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 struc- ture 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 work has been presented to a conference [27] and as a poster to a workshop [36]. It is a joint work with Gaétan Loridant from Arcelor-Mittal.

6.9. Model-based clustering for multivariate functional data

Participants: Julien Jacques, Cristian Preda.

We developed in [19] an extension of the model-based clustering algorithm for univariate functional data proposed in [20], [23], [11] to the case of multivariate functional data. For this, multivariate functional principal components analysis is defined and a parametric mixture model is proposed and estimated by an EM-like algorithm. Results on simulated and real datasets have shown the efficiency of the proposed method.

6.10. A method to combine combinatorial optimization and statistics to mine high-throughput genotyping data

Participants: Julie Hamon, Julien Jacques, Clarisse Dhaenens.

In the context of genomic analysis (collaboration with Genes Diffusion), dealing with high-throughput genotyping data, the objective of our study is to select a subset of SNPs (single nucleotide polymorphisms) explaining a trait of interest. We propose in [33] and [32] a method combining combinatorial optimization and statistics to extract a subset of interesting SNPs. The combinatorial part aims at exploring in a efficient way the large search space induced by the large number of possible subsets and statistics are used to evaluate the selection. We propose a first method based on an ILS (iterated local search) and using a regression. Three criteria used to evaluate the quality of the regression are compared. One of them (the k-fold validation) shows better performance. We also compare this approach to classical statistical approaches on simulated datasets. Results are promising as the proposed approach outperforms most of these statistical approaches.

6.11. Wavelet based clustering using mixed effects functional models

Participant: Guillemette Marot.

Curve clustering in the presence of inter-individual variability has longly been studied, especially using splines to account for functional random effects. However splines are not appropriate when dealing with high-dimensional data and can not be used to model irregular curves such as peak-like data. We propose a wavelet based clustering procedure ([6]) and apply it to high dimensional data. We suggest a dimension reduction step based on wavelet thresholding adapted to multiple curves and using an appropriate structure for the random effect variance, we ensure that both fixed and random effects lie in the same functional space even when dealing with irregular functions that belong to Besov spaces. In the wavelet domain, our model resumes to a linear mixed-effects model that can be used for a model-based clustering algorithm and for which we develop an EM- algorithm for maximum likelihood estimation. An R package curvclust implementing this procedure has been posted this year to the CRAN, the official website of the R software.

6.12. Comparison of normalisation procedures in RNA-sequencing before differential analysis

Participant: Guillemette Marot.

The continuing technical improvements and decreasing cost of next-generation sequencing technologies have made RNA sequencing (RNA-seq) a popular choice for gene expression studies. Several methods for the normalization of RNAseq data (removal of errors due to the small number of samples, corrections for sequence composition) have been proposed in recent years. With the Statomique Consortium, we have compared seven normalisation methods, discarded two out of them (although still widely used). We give practical recommendations on the appropriate normalization method to be used and its impact on the differential analysis of RNA-seq data in the paper ([14]).

6.13. Change point detection algorithm

Participant: Alain Célisse.

We develop a new change-point detection algorithm where focus is given to detect changes in the whole distribution of data. This challenging problem is addressed by use of kernels which enable us to deal with non-vectorial data of aby type (graphs, DNA sequences, etc). A preprint has been submitted ([46]).

6.14. Cross validation algorithms

Participant: Alain Célisse.

The performance of Cross-validation (CV) algorithms are assessed for estimating the risk as well as for model selection. Whereas optimality of leave-one-out (LOO) cross-validation is proved for risk estimation, it is no longer the case for model selection. In the latter setup, conditions are derived that lead to optimality for leave-p-cross-validation (LPO) when p is larger than 1. See for details [47].

6.15. Stochastic Block Model

Participant: Alain Célisse.

The convergence of maximum likelihood and variational estimators in a random graph model called Stochastic Block model is addressed. To the best of our knowledge, these are the first results providing consistency for maximum likelihood and variational estimators in that model. See [5].

6.16. Approximations for scan statistics.

Participants: Alexandru Amarioarei, Cristian Preda.

Accurate approximations for the distribution of extremes of 1-dependent stationary sequences are developed (see [38]). Viewed as maximum of some particular sequence of 1-dependent random variables, we provide sharp error bounds and approximations for the distribution of the three-dimensional scan statistics (see [39]). The Binomial and Poisson models are considered.

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 semi-supervised 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. Gene Diffusion

Participants: Julien Jacques, Julie Hamon.

Subject : *Data analysis from high throughput technologies: Synergy between statistics and combinatorial optimization.*

With the development of new technologies such as high-throughput genotyping and sequencing, data analysis needs to be improved. Genes Diffusion is specialized in animals studies, for which we can read genomics information on around 800 000 markers and we have more and more subjects. The aim of the PhD is to find new methods combining combinatorial optimization and statistics methods in order to characterize the best subjects according to quantitative criteria. A PhD CIFRE grant started on 2010 and it is a joined work with Clarisse Dhaenens (Inria/DOLPHIN).

7.3. ASEL & CRESGE

Participants: Cristian Preda, Michael Genin.

Subject : *Incidence of lymphoms in Nord-Pas-de-Calais, Annual Estimates and study of the evolution over the period 2001-2005.* It is a contract with ASEL (Association Septentrionale pour l'Etude de Lymphomes) and CRESGE (Centre de Recherches Economiques Sociologiques et de Gestion) from Lille. This project of 6000 euros started on September 1st 2012 and ends on Mai 1st 2013.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Institut de Biologie de Lille, Génomique et Maladies Métaboliques lab

Participants: Christophe Biernacki, Julien Jacques, Loic Yengo.

8.1.2. Industrial Studies Center, Arcelor-Mittal

Participants: Clément Thery, Christophe Biernacki.

8.1.3. Gene Diffusion

Participants: Julien Jacques, Julie Hamon.

8.1.4. Institut Pasteur Lille and Institut de Biologie de Lille

Participant: Guillemette Marot.

- Team "Etudes Transcriptomiques et Génomiques Appliquées"n (D. Hot).
- Team "Peste et Yersinia pestis", (F. Sebbane).
- team "Unité d'approches fonctionnelle et structurale des cancers", O. Pluquet.

8.1.5. Université de Lille 2

Participant: Guillemette Marot.

Plate-forme de génomique fonctionnelle et Structurale, (M. Figeac)

8.1.6. CHRU Lille

Participant: Guillemette Marot.

Centre de Biologie Pathologie, Laboratoire d'Hématologie, (C. Preudhomme)

8.1.7. ASEL and CRESGE

Participant: Cristian Preda.

ASEL (Association Septentrionale pour l'Etude de Lymphomes) and CRESGE (Centre de Recherches Economiques Sociologiques et de Gestion) from Lille

8.2. National Initiatives

8.2.1. StatLearn'12

Christophe Biernacki, Alain Ceélisse, Serge Iovleff and Julien Jacques co-organized with Charles Bouveyron (University Paris 1, SAMM) a workshop on "Challenging problems in Statistical Learning", StatLearn'12, in April 2012 in Lille (http://www.inria.fr/en/centre/lille/calendar/workshop-statlearn-12). There were about 80 applicants, 12 one-hour invited talk organized in four sessions: Statistical learning and vizualization, Statistical learning in high dimension, Statistical learning and structured data, New and future problems in statistical learning.

8.2.2. StatOmique

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

8.3. European Initiatives

8.3.1. University of Granada, Department of Statistics and Operational Research

Participant: Cristian Preda.

Collaboration with Professor Ana Aguilera : teaching at Master and Doctoral level, joint research, ERASMUS mobility and conference organization.

8.4. International Research Visitors

8.4.1. Nanyang Technology University of Singapore

Participant: Cristian Preda.

Collaboration with Professor Lian Heng on functional regression models : joint research.

Cristian Preda was invited at NTU from December 3th to December 15th 2012.

9. Dissemination

9.1. Scientific Animation

Since '12, C. Biernacki participates to the international group "IFCS Committee on Initiative to Stimulate Benchmarking in Classification Research". 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/.

Guillemette Marot organizes, in the context of the PPF bioinfo Lille 1, two scientific meetings:

- Phylogénie (35 people), Lille 1, june 2012, http://www.lifl.fr/~touzet/PPF/phylogenie12.html
- Analyse bioinformatique des données de métagénomique (80 people expected), Pasteur Lille, december 2012, http://www.lifl.fr/~touzet/PPF/metagenomique12.html

9.1.1. Editorial activities

C. Biernacki belongs to the program committe of "Extraction et gestion des connaissances" in 2012 and 2013 and to the program comity of "Journées Françaises de Statistique" in 2013. He organised a special session "clustering of mixed data" in the conference SFdS 2012 in Brussels. Since '10, he is an Associate Editor of the journal "Case Studies in Business, Industry and Government Statistics" (CSBIGS) http://legacy.bentley.edu/csbigs/.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence : Cristian Preda, Probabilités, 36h, L3, École Polytechnique Universitaire de Lille, U. Lille 1, France.

Licence : Julien Jacques, Statistique Inférentielle, 50h, L3, École Polytechnique Universitaire de Lille, U. Lille 1,France

Licence : Guillemette Marot, Biostatistique, 18h, L1, U. Lille 2, France

Licence : Alain Célisse, Algèbre, 80h, L2, U. Lille 1, France

Licence : Alain Célisse, Mathématiques pour l'Informatique, 122h, L1, U. Lille 1, France

Licence : Serge Iovleff, Analysis, 24h, U. Lille 1, France.

Licence : Serge Iovleff, Probability and Statistics, 24h, U. Lille 1, France.

Licence : Serge Iovleff, Discrete mathematics and algebra, 72h, U. Lille 1, France

Licence : Serge Iovleff, Graphes and Languages, 80h, U. Lille 1, France

Licence : Serge Iovleff, Algebra, Geometry and Arithmetic, 32h, U. Lille 1, France

Licence : Serge Iovleff, Options of Mathematics, 48h, U. Lille 1, France

Licence : Vincent Vandewalle, Linear algebra, 91h, Simulation Techniques, 16h, Descriptive statistics, 62h, Probabilities, 44h, L1 , U. Lille 2, France

Licence : Vincent Vandewalle, Analysis, 24h, Project management, 9h, L2, U. Lille 2, France

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

Licence : Matthieu Marbac-Lourdelle, Data analysis, 48h, Institut Superieur d'Agriculture, U. catholique de Lille, France.

Master : Cristian Preda, Statistique Exploratoire, 40h, M1, École Polytechnique Universitaire de Lille, U. Lille 1,France.

Master : Cristian Preda, Functional Data Analysis, 18h, M2, U. Lille 1, France.

Master : Julien Jacques, Statistique Exploratoire, 40h, M1, École Polytechnique Universitaire de Lille, U. Lille 1,France.

Master : Julien Jacques, Modélisation Statistique, 30h, M1, École Polytechnique Universitaire de Lille, U. Lille 1,France

Master : Julien Jacques, Séries Temporelles, 25h, M2, École Polytechnique Universitaire de Lille, U. Lille 1, France

Master : Guillemette Marot, Biostatistique, 48h, M1, U. Lille 2, France

Master : Alain Célisse, Statistique Fondamentale, 45h, M2, U. Lille 1, France

Doctorat : Cristian Preda, Functional Data Analysis, 10h, M2, Department of Statistics, University of Granada, Spain.

Christophe Biernacki was in a one year delegation at Inria with no teaching. From '05, he is the 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/le-departement-stid.html

9.2.2. Supervision

HdR : Julien Jacques, Contribution to statistical learning of complex data using generative models, Université Lille 1, November 28, 2012.

PhD in progress : Alexandru Amarioarei, Statistics, Scan statistics and applications, started in 2010, Cristian Preda supervisor

PhD in progress : Michael Genin, Statistics, Scan statistics and epidemiology, started in 2010, Cristian Preda and Alain Duhamel (CEREM, U. Lille 2) supervisors

PhD in progress : Julie Hamon, Analysis of data from high throughput genotyping: cooperation between statistics and combinatorial optimization, started in 2010, Julien Jacques and Clarisse Dhaenens (DOLPHIN Inria Lille team-project) supervisor

PhD in progress : Loïc Yengo, Simultaneous Variables Clustering and Selection in Regression Models, started in 2010, Christophe Biernacki and Julien Jacques supervisors

PhD in progress : Clément Thery, Classification supervisée ou semi-supervisée des bases de grande dimension, avec variables qualitatives et quantitatives, started in 2011, Christophe Biernacki supervisor

PhD in progress : Matthieu Marbac-Lourdelle, Generatives models taking into account the correlation between variables, started in 2011, Christophe Biernacki and Vincent Vandewalle supervisors

10. Bibliography

Major publications by the team in recent years

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

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

[3] 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, p. 2991—3002.

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

[5] A. Célisse, J.-J. Daudin, L. Pierre. *Consistency of maximum likelihood and varitional estimators in stochastic block moddel*, in "Electronic Journal of Statistics", 2012, p. 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", 2012, to appear.

[7] M. Guedj, A. Célisse, G. Nuel. *kerfdr: A semi-parametric kernel-based approach to local FDR estimations*, in "BMC Bioinformatics", 2009, vol. 84, n^o 10, (electronic).

[8] 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, p. 749–766.

[9] J. JACQUES, C. PREDA. *Funclust : a curves clustering method using functional random variable density approximation*, in "Neurocomputing", 2012, to appear.

[10] A. LOURME, C. BIERNACKI. *Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins*, in "Computational Statistics", 2012, to appear.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] J. JACQUES. *Contribution to statistical learning of complex data using generative models*, University Lille 1, 2012.

Articles in International Peer-Reviewed Journals

[12] A. CÉLISSE, J.-J. DAUDIN, L. PIERRE. *Consistency of maximum likelihood and varitional estimators in stochastic block moddel*, in "Electronic Journal of Statistics", 2012, p. 1847–1899, http://projecteuclid.org/handle/euclid.ejs.

[13] V. DAMIEN, H. ISABELLE, D. SÉVERINE, D. SÉBASTIEN, G. MAROT, D. OLIVIER, G. GUY, H. PATRICE, P. ANDREW, C. GILLES, G. BÉNÉDICTE. *Pre- and Post-Partum Mild Underfeeding Influences Gene Expression in the Reproductive Tract of Cyclic Dairy Cows*, in "Reproduction in Domestic Animals", November 2012, http://dx.doi.org/10.1111/rda.12113.

[14] M.-A. DILLIES, A. RAU, A. JULIE, M. JEANMOUGIN, N. SERVANT, C. KEIME, G. MAROT, D. CASTEL, E. JORDI, G. GUERNEC, J. BERND, L. JOUNEAU, D. LALOE, C. LE GALL, B. SCHAEFFER, S. LE CROM, M. GUEDJ, F. JAFFRÉZIC. *A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis*, in "Briefings in Bioinformatics", September 2012.

[15] 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.

[16] J. JACQUES, C. PREDA. *Funclust : a curves clustering method using functional random variable density approximation*, in "Neurocomputing", 2012, to appear.

[17] A. LOURME, C. BIERNACKI. *Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins*, in "Computational Statistics", 2012, to appear.

International Conferences with Proceedings

[18] E. EIROLA, A. LENDASSE, V. VANDEWALLE, C. BIERNACKI. *Mixture of Gaussians for Distances Estimations with Missing Data*, in "Workshop New Challenges in Neural Computation", March 2012, Machine Learning Reports 03/2012.

[19] J. JACQUES, C. PREDA. *Clustering multivariate functional data*, in "COMPSTAT'12", Limassol, Chypre, August 2012.

[20] J. JACQUES, C. PREDA. *Curves clustering with approximation of the density of functional random variables*, in "20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'12)", Bruges, Belgium, April 2012.

National Conferences with Proceeding

[21] C. Biernacki, J. Jacques. *Modèle génératif pour données ordinales*, in "44e Journées de Statistique organisée par la Société Française de Statistique", Bruxelles, 2012.

[22] J. Jacques. *Classification automatique de données hétérogènes*, in "44e Journées de Statistique organisée par la Société Française de Statistique", Bruxelles, 2012.

[23] J. Jacques, C. Preda. *Functional data clustering using density approximation*, in "44e Journées de Statistique organisée par la Société Française de Statistique", Bruxelles, 2012.

[24] R. Lebret, S. Iovleff, C. Biernacki, J. Jacques, C. Preda, M. Alun, D. Olivier. *Genetic epistasis analysis using taxonomy3*, in "2nd International BIO-SI Workshop", Rennes, France, October 2012.

[25] R. Lebret, S. Iovleff, C. Biernacki, J. Jacques, C. Preda, M. Alun, D. Olivier. *Rapid multivariate analysis of 269 hapmap subjects and 1 million snps using taxonomy3*, in "Cold Spring Harbor/Wellcome Trust meeting on Pharmacogenomics", Hinxton, UK, September 2012.

[26] M. Marbac, C. Biernacki, V. Vandewalle. *Modèle de mélange pour classifier des données qualitatives conditionnellement corrélées*, in "44e Journées de Statistique organisée par la Société Française de Statistique", Bruxelles, 2012.

[27] C. Thery, C. Biernacki, G. Loridant. *Décorrélation de variables en régression linéaire par modèles de sous-régressions*, in "44e Journées de Statistique organisée par la Société Française de Statistique", Bruxelles, 2012.

[28] L. Yengo, J. Jacques, C. Biernacki. *Classification et sélection de variables en régression*, in "44e Journées de Statistique organisée par la Société Française de Statistique", Bruxelles, 2012.

Conferences without Proceedings

[29] A. Amarioarei, C. Preda. *Approximation for the distribution of three-dimensional scan statistics*, in "International Workshop on Applied Probability", Jerusalem, Israel, June 8-15 2012.

[30] P. Bhatia, S. Iovleff, G. Govaert. *Sofware for Co-clustering of Binary, Contingency and Continuous Data-sets*, in "Workshop on Challenging problems in statistical learning", Lille, France, April 2012, Poster presentation.

[31] C. Biernacki, A. Lourme. *Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection*, in "MBC 2 - Workshop on Model Based Clustering and Classification", Catania, Italy, September 6-7 2012.

[32] J. Hamon, C. Dhaenens, J. Jacques, G. Even. *Coopération entre optimisation combinatoire et statistique pour la sélection animale*, in "13e congrès annuel de la Société française de Recherche Opérationnelle et d'Aide à la Décision", Angers, France, April 2012.

[33] J. Hamon, C. Dhaenens, J. Jacques, G. Even. *Feature selection for high dimensional regression using local search and statistical criteria*, in "4th International Conference on Metaheuristics and Nature Inspired Computing, META'2012", Sousse, Tunisia, October 2012.

[34] R. LEBRET, S. IOVLEFF, F. LANGROGNET. *Rmixmod: A MIXture MODelling R Package*, in "Statlearn'12 - Workshop on Challenging problems in statistical learning", Lille, France, April 2012.

[35] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Model-based clustering for conditionally correlated categorical data*, in "Statlearn'12, Workshop on "Challenging problems in Statistical Learning"", Lille, France, April 2012.

[36] C. THERY, C. BIERNACKI, G. LORIDANT. *Decorrelating variables in high dimension for linear regression*, in "Statlearn'12, Workshop on "Challenging problems in Statistical Learning"", Lille, France, April 5-6 2012.

Scientific Books (or Scientific Book chapters)

[37] F. BENINEL, C. BIERNACKI, C. BOUVEYRON, J. JACQUES, A. LOURME. *Parametric link models for knowledge transfer in statistical learning*, Knowledge Transfer: Practices, Types and Chalanges, Nova Publishers, 2012.

Research Reports

[38] A. AMARIOAREI. *Approximation for the distribution of extremes of 1-dependent stationary sequences of random variables*, preprint, 2012.

[39] A. AMARIOAREI, C. PREDA. *Approximation for the distribution of three-dimensional scan statistics*, preprint, 2012.

[40] P. BHATIA, S. IOVLEFF, G. GOVAERT. *blockcluster: An R Package C++ library for Latent block models : Theory, usage and application.*, preprint, 2012.

[41] C. BIERNACKI, A. LOURME. *Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection*, Inria, 2012, n^o 7932.

[42] C. BIERNACKI, A. LOURME. *Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection*, Inria, April 2012, n^o RR-7932, 21, http://hal.inria.fr/hal-00688250.

[43] J. JACQUES, C. BIERNACKI. *Model-based clustering for multivariate partial ranking data*, Inria, October 2012, n^o RR-8113, 25, http://hal.inria.fr/hal-00743384.

[44] 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*, preprint, 2012.

[45] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Model-based clustering for conditionally correlated categorical data*, preprint, 2012.

Other Publications

[46] S. ARLOT, A. CÉLISSE, Z. HARCHAOUI. *Kernel change-point detection*, 2012, http://hal.inria.fr/hal-00671174.

[47] A. CÉLISSE. *Optimal cross-validation in density estimation*, 2012.

[48] S. IOVLEFF. *Probabilistic Auto-Associative Models and Semi-Linear PCA*, 2012, http://hal.inria.fr/hal-00734070.

[49] J. JACQUES, C. PREDA. *Model-based clustering for multivariate functional data*, 2012, http://hal.inria.fr/hal-00713334.