Activity Report 2011

Team MODAL

MOdel for Data Analysis and Learning

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, Learning and Statistical Methods

Table of contents

Team MODAL

Keywords: Statistical Learning, Data Analysis, Classification, Visualization

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, chair INRIA]
Vincent Vandewalle [Associate Professor at U. Lille 2, EA 2694]

Technical Staff

Parmeet Bhatia [ADT grant]
Remi Lebret [U. de Technologie de Compiègne, UMR 6599]

PhD Students

Alexandru Amariorei [MESR grant]
Michael Genin [MESR grant]
Julie Hamon [CIFRE grant]
Matthieu Marbac-Lourdelle [DGA-INRIA grant]
Alexandre Lourme [U. Pau]
Clément Thery [CIFRE grant]
Loïc 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

Since November 2011, the team started the development of the co-clustering module in the MIXMOD software, allowing to process efficient and parsimonious generative models on huge data sets (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 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 application areas are considered as *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

5.1. MIXMOD

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

MIXMOD (MIXture MODelling) is the core software of the MODAL team for two reasons. First, MIXMOD concerns main topics of MODAL since it is devoted to model-based supervised, unsupervised and semi-supervised classification for various data situations. Second, MIXMOD is now a well-distributed software since over 250 downloads/month are recorded for several years. Consequently, MIXMOD will be the main software for diffusing future methodological advances of the MODAL team.

MIXMOD is written in C++ (more than 10 000 lines), currently interfaced with Scilab and Matlab and distributed under GNU General Public License. An interface between MIXMOD and R is being to be developed by Rémi Lebret and will be soon available (during 2012).

Several other institutions partiticipate 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 leads also to some international publications[1].

In addition, an INRIA ADT grant (Parmeet Bhatia) will also develop co-clustering models for continuous, binary and discrete data. It is a strategic development for MIXMOD since offering the ability to structure very large data tables both in lines and columns for different data types. In particular, it opens wide potential applications in biology, marketing, *etc.*

Serge Iovleff is the main supervisor of software engineers who are recruited for all the previously described tasks. More information about MIXMOD can be easily found on its web page http://www.mixmod.org/.

5.2. AAM

Participant: Serge Iovleff.

The AAM program is a R library implementing Auto-Associative models. Thus it could with few work transformed into a R package. As the AAM is a statistical model, the R language was well-suited for a diffusion inside the scientific community. It is a prototype for testing the AAM models against other kind of non-linear PCA models.

The first release was a scilab program written by Serge Iovleff and Stéphane Girard. It was rewriten in January 2009 and the code is now faster and produces enhanced graphics. The 2009 release is the result of a conjoint work of Serge Iovleff and a M1 internship of the ENS.

More information on the web site http://www.iut-info.univ-lille1.fr/~iovleff/softwares/

5.3. Kerfdr

Participant: Alain Célisse.

Computation of the local FDR: R package for biostatisticians allowing to estimate FDR and local FDR by kernel density estimation. This package allows also to deal with truncated data and to take into account supervision. More information on the website http://cran.r-project.org/web/packages/kerfdr/

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 this 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.

[1] C. Biernacki, G. Celeux, G. Govaert and F. Langrognet, *Model-Based Cluster and Discriminant Analysis with the* MIXMOD *Software*, Computational Statistics and Data Analysis, Vol. 52, no 2, 587–600, 2006

First versions have been 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 publications citing the software[2].

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

5.5. 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... For instance, some part of the project will be integrated to the co-cluster project (in the MIXMOD software, see Section 5.1) actually developed by Parmeet Bathia.

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

5.6. 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 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. She received some help from Anne de la Foye (INRA, Clermont-Ferrand) to correct the bugs in the first versions.

First versions have been 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 publications citing the software[3].

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

6. New Results

6.1. Intermediate dependency generative models

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

[2] G. Marot,J.-L. Foulley, C.-D. Mayer, F. Jaffrézic, *Moderated effect size and p-value combinations for microarray meta-analyses*, in "Bioinformatics", 2009, vol. 25, no 20, p. 2692—2699.

[3] F. Jaffrézic, G. Marot, S. Degrelle, I. Hue, J.-L. Foulley. A structural mixed model for variances in differential gene expression studies, in "Genetical Research", 2007, vol. 89, no 1, p. 19—25.

Defining generative models for dealing with possibly correlated categorical variables is at the core of the MODAL activity. We start by noticing that it is straightforward to build a full independent distribution $\mathring{p}$ and also a full dependent one $\acute{p}$ in the categorical situation. However, both are usually too crude for modelizing most of real situations.

Our idea is to combine both extreme distributions $\mathring{p}$ and $\acute{p}$ in order to obtain a new distribution called $\widetilde{p}$ *(i)* which is an intermediate dependent situation between full independence and full dependence and *(ii)* which is not degenerate. As a consequence, $\widetilde{p}$ is a meaningful distribution because its particular “positioning” between $\mathring{p}$ and $\acute{p}$ directly models and reveals strength of dependency between variables.

In addition, since both $\mathring{p}$ and $\acute{p}$ are easily available for most variables types, we expect to be able to design a distribution $\widetilde{p}$ for most variables types, and not also the categorical ones.

A PhD thesis started on October’11 on this topic in continuation of the Master’s thesis of Matthieu Marbac-Lourdelle [37].

6.2. Transfer learning in model-based clustering

Participants: Christophe Biernacki, Alexandre Lourme.

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 ([15] and a paper in revision). In the general context (categorical or heterogeneous variables), we propose to use alternatively an entropic link between populations [17]. All these works are related to the Lourme’s PhD thesis [11].

A chapter of book about transfer learning (including clustering, classification and regression) is currently submitted for publication (joint work with Julien Jacques and Alexandre Lourme).

6.3. Block regression for variable clustering: Application to genetic data

Participants: Christophe Biernacki, Julien Jacques, Loïc Yengo.

Genome Wide Association (GWA) studies have proved the implication of numerous single nucleotides polymorphisms (SNP) in the etiology of common diseases. Nevertheless, only a small part of the expected heritability of those diseases is explained by the most significantly associated SNPs. Many researches that have been lately investigating this missing heritability have considered interactions between genes and/or environmental factors as a plausible and promising explanation. Considering all if not a large number (hundreds of thousands) of variants altogether stresses the problem of the high dimensionality that most regression-based methods cannot afford. To solve this issue one either reduces the number of variants to be analyzed (shrinkage approaches) or groups them according to a certain similarity. We introduce here a regression model that simultaneously clusterizes the variants sharing close effect size while selecting the most informative clusters. The estimation of the model parameters is proposed by maximizing the likelihood. The challenges of this research rely on finding efficient algorithms for the clustering part while studying the consistency of our estimators for which the classical asymptotic theory does not apply [33], [40].

6.4. Label switching in mixtures

Participants: Christophe Biernacki, Vincent Vandewalle.

During the last fifteen years there has been an increasing interest for using Bayesian methods in mixtures models. However, one of the principal issues of these methods is the non-identifiability of components caused by symmetric prior (whatever be the kind of variables), which makes the Gibbs outputs useless for inference; this problem is known as label switching. We propose to condition the posterior distribution by a particular numbering, not on the parameter as it is usually done, but rather on a latent partition, for which the posterior distributions are not any more strictly invariant up to a renumbering of the partition [26], [19], [32]. The importance of this asymmetry depends on the choice of partition space cutting. The challenge we address is to choose a particular cutting which is justified and also easy to compute. The idea is to use some properties of the (unavailable) *completed* posterior distribution.

6.5. Degeneracy in 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. Currently, such results are proved in the univariate case [34]. The challenge is now not only to extend them in the multivariate situation but also to complete these theoretical results with some practical strategies for properly avoiding degeneracy in softwares devoted to such mixture estimations.

This is a joined work with Gwënaelle Castellan.

6.6. 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 [23] 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 is under building and should be posted to the CRAN, the official website of the R software, before Dec. 2011. An article has been submitted once to Biometrics and received good reports. This paper should also be submitted again to Biometrics once curvclust is on the CRAN.

6.7. 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 in recent years. Because the data collected from such studies differ considerably from those measured using microarray technology, the statistical tools used for analysis must be adapted accordingly. In particular, 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. First results are given in [28].

6.8. Comparison of peak finding methods applied to tiling array experiments

Participant: Guillemette Marot.

Scan statistics are widely used to detect peaks in tiling array experiments. An extensive analysis study of real biological data is being performed with Florent Sebbane and David Hot teams (Institut Pasteur, Lille) for the study of the Yersinia Pestis bacteria in order to find new small RNAs. First results have been presented in [31]. Given a signal composed of intensities ordered along the genome, the statistical problem is to detect peaks, taking into account the irregular design of the chips, which the biologists had chosen a few years ago. A master student (D. Thuillier) has compared different normalisation methods during a 6 months internship and improved the first analysis results presented in [31]. We also propose a local score procedure, which seems promising according to first biological results obtained. The next step is to work with Alain Célisse in order to choose a generative model on the normalised data which would enable to give appropriate initial values to the local score procedure and associate p-values to local scores.

6.9. Model-based clustering for functional data

Participants: Julien Jacques, Cristian Preda.

Two procedures for clustering functional data have been developed.

The first one, published in [14], is based on a functional latent mixture model which fits the functional data in group-specific functional subspaces. By constraining model parameters within and between groups, a family of parsimonious models is exhibited which allows to fit onto various situations. An estimation procedure based on the EM algorithm is proposed for estimating both the model parameters and the group-specific functional subspaces. Experiments on real-world datasets show that the proposed approach performs better or similarly than classical clustering methods while providing useful interpretations of the groups.

The second procedure, currently submitted, is a model-based clustering procedure, defined on the basis of an approximation of the density of functional random variables [36]. As previously, the EM algorithm is used for parameter estimation and the maximum a posteriori rule provides the clusters. Simulation study and real data application illustrate the interest of this methodology.

6.10. Generative models and random graphs

Participant: Alain Célisse.

The aim is to study consistency of variational and maximum likelihood estimates built from a particular generative model of random graph where independence between the ridges of the graph is not assumed. These results are established from concentration inequalities. They have a great practical interest since they justify *a posteriori* intensive use of variational methods in this context.

It is a joint collaboration with Jean-Jacques Daudin and a paper is submitted [35].

6.11. Resampling and learning

Participant: Alain Célisse.

This aim is to study the k nearest neighbors algorithm in binary classification in two different cases: Passive and active learning. The choice of k is addressed by cross-validation (resampling). In particular, we try to discover the influence of the cutting parameter on which depends the cross-validation with the retained k value.

It is a joined work with Tristan Mary-Huard [16].

7. Contracts and Grants with Industry

7.1. Genes Diffusion

Participants: Julien Jacques, Julie Hamon.

“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 (DOLPHIN).

7.2. Natural Security

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

“Statistical modeling and simulation for card payment at medium distance.”

As part of the "Payment of a hand gesture", the Natural Security company uses a technology at medium distance based on biometrics to authenticate owner and to allow transaction with no card payment manipulation while limiting the risk fraud. Depending to the context of use (frequency of transactions) a theoretical expertise is needed to assess the viability of the system in term of probability of collision or wrong authentication. This collaboration has led to two contracts in 2011, 6 k€ each and about two weeks long each.

7.3. Arcelor-Mittal

Participants: Christophe Biernacki, Clément Thery.

“Supervised and semi-supervised classification on large data bases mixing qualitative and quantitative variables.”

Arcelor-Mittal is faced with 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 number of variables, (2) heterogeneous variables, typically categorical and continuous variables and (3) potentially highly correlated variables. A PhD CIFRE grant started on May 2011.

7.4. ASEL & CRESGE

Participant: Cristian Preda.

“Incidence of lymphoma 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 6 k€ starts on December 1st 2011 and ends on September 2st 2012.

8. Partnerships and Cooperations

8.1. Regional Initiatives

Christophe Biernacki and Julien Jacques:

- Institut de Biologie de Lille, laboratory Génomique et Maladies Métaboliques, L. Yengo

Christophe Biernacki:

- Industrial studies, Arcelor-Mittal, C. Théry

Julien Jacques:

- Genes Diffusion, J. Hamon

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

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

- Alain Célisse co-organized a workshop on Random Graphs in Lille on April '11 http://math.univ-lille1.fr/~tran/journeesgraphesaleatoires.html. There were about 50 applicants, about 12 one-hour talks, one lecture (4 hours and a half) on probabilitic aspects on random graphs by Remco von der Hofstad
- Guillemette Marot belongs to the StatOmique working group http://vim-iip.jouy.inra.fr:8080/statomique/

8.3. European Initiatives

8.3.1. Major European Organizations with which you have followed Collaborations

Partner 1: University of Granada, Department of Statistics (Spain)

Collaboration with Professor Ana Aguilera in the field of Functional Data Analysis. Form of collaboration: joint paper[4], ERASMUS mobility, conference organization.

Partner 2: Luxembourg School of Finance (Luxembourg)

Collaboration with Professor Jang Schiltz for time-series prediction using functional data. Form of collaboration : joint paper [25], mobility projects research.

9. Dissemination

9.1. Animation of the scientific community

9.1.1. Editorial responsibilities

C. Biernacki belongs to the scientific committee of “Model mixtures and learning” in SFdS'11 (Gammarth, Tunisia) and to the program comity of "Extraction et gestion des connaissances" in EGC'12 (Bordeaux, France). Since '10, he is an Associate Editor of the journal “Case Studies in Business, Industry and Government Statistics” (CSBIGS) http://legacy.bentley.edu/csbigs/.

[4] A. Aguilera, M. Escabias, C. Preda, G. Saporta, *Using basis expansion for estimating functional PLS regression: application with chemometric data*, in "Chemometrics and Intelligent Laboratory Systems", Vol. 104, no 2, p. 289–305, 2010

9.1.2. Invited conferences

- C. Biernacki and V. Vandewalle are invited speakers to one conference [19]
- C. Preda is an invited speaker in '11 to three conferences [18], [20], [21]

9.1.3. Scientific animation

- Since '09, C. Biernacki is a treasurer 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/. In '11, he reviewed 3 PhD theses.
- Cristian Preda:
  - organized a session of applied statistics for the Statistics and Probability Society of Romania (Bucarest, April 2011)
  - was Scientific Supervisor for the statistical methodology developed in the PSIP FP7 European project http://psip-project.eu
  - performed research conferences and teaching on statistics at the University of Granada, University of Luxembourg and University of Bucharest
  - organized several conferences for the Seminar of Statistics and Informatics of the Faculty of Medicine, University Lille 2.
- Guillemette Marot organizes, in the context of the PPF bioinfo Lille 1, two scientific meetings:
  - Fouille de texte pour la biologie, Sept. 2011, http://www.lifl.fr/~touzet/PPF/fouilletexte11.html
  - Analyse bioinformatique des données NGS, Dec. 2011, http://www.lifl.fr/bonsai/seqbio2011/ngs11.html

9.2. Teaching

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

Master: Mathematical statistics, 60h, coaching project, 10h, M1, U. Lille 1, France

Master: Data analysis, 97.5h, Analysis of variance and experimental design, 22.5h, coaching internship, 20h, M2, U. Lille 2, France

Alain Célisse:

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

DUT: Mathématiques pour l'Informatique, 122h, L1, U. Lille 1, France

DUT: Algèbre, 80h, L2, U. Lille 1, France

Serge Iovleff:

DUT: Discrete mathematics, 72h, Modelization, 88h, Algebra & Geometry, 32h, Probability & statistics & analysis 64h, L1, U. Lille 1, France

Julien Jacques:

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

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

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

Guillemette Marot:

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

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

Cristian Preda:

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

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

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

Master: Functional Data Analysis, 10h, M2, Department of Statistics, University of Granada, Spain

Vincent Vandewalle:

DUT STID: Linear algebra, 93h, Simulation Technics, 31.5h, Descriptives statistics, 36h, Basic mathematics, 12h, Probabilities, 108h, L1 , U. Lille 2, France

DUT STID: Analysis, 20h, L2, U. Lille 2, France

PhD: Alexandre Lourme, Contribution à la Classification par Modèles de Mélange & Classification Simultanée d'Echantillons d'Origines Multiples, U. Lille 1, June'11, Christophe Biernacki supervisor

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

PhD in progress : Michael Genin, Statistics, Scan statitics 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. Bouveyron, J. Jacques. *Adaptive linear models for regression: Improving prediction when population has changed*, in "Pattern Recognition Letters", 2010, vol. 31, n^o 14, p. 2237–2247.

[5] S. Girard, S. Iovleff. *Auto-associative models, nonlinear Principal component analysis, manifolds and projection pursuit*, Principal Manifolds for Data Visualisation and Dimension Reduction, In A. Gorban et al, editors, 2007, vol. 8, LNCSE, Springer-Verlag.

[6] 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).

[7] J. Jacques, C. Biernacki. *Extension of model-based classification for binary data when training and test populations differ*, in "Journal of Applied Statistics", 2010, vol. 37, n^o 5, p. 749–766.

[8] G. Marot, J.-L. Foulley, C.-D. Mayer, F. Jaffrézic. *Moderated effect size and p-value combinations for microarray meta-analyses*, in "Bioinformatics", 2009, vol. 25, n^o 20, p. 2692–2699.

[9] C. Preda. *Regression models for functional data by reproducing kernel Hilbert spaces methods*, in "Journal of Statistical Planning and Inference", 2007, vol. 37, n^o 3, p. 829–840.

[10] C. Preda, G. Saporta, C. Lévéder. *PLS classification for functional data*, in "Computational Statistics", 2007, vol. 22, p. 223–235.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] A. Lourme. *Contribution à la Classification par Modèles de Mélange & Classification Simultanée d'Echantillons d'Origines Multiples*, University Lille 1, 2011.

Articles in International Peer-Reviewed Journal

[12] S. Arlot, A. Célisse. *Segmentation in the mean of heteroscedastic data by cross-validation*, in "Statistics and Computing", 2011, vol. 21, n^o 4, p. 613–632.

[13] C. Bouveyron, P. Gaubert, J. Jacques. *Adaptive models in regression for modeling and understanding evolving populations*, in "Case Studies in Business, Industry and Government Statistics (CSBIGS)", 2011, vol. 4, n^o 2.

[14] C. Bouveyron, J. Jacques. *Model-based Clustering of Time Series in Group-specific Functional Subspaces*, in "Advances in Data Analysis and Classification", December 2011, vol. 5, n^o 4, p. 281–300, http://hal.inria.fr/hal-00559561/en.

[15] A. Lourme, C. Biernacki. *Simultaneous t-Model-Based Clustering for Data Differing over Time Period: Application for Understanding Companies Financial Health*, in "Case Studies in Business, Industry and Government Statistics (CSBIGS)", 2011, vol. 4, n^o 2.

Articles in National Peer-Reviewed Journal

[16] A. Célisse, T. Mary-Huard. *Exact Cross-Validation for kNN and applications to passive and active learning in classification*, in "Journal de la Société Française de Statistique", 2011.

[17] A. LOURME, C. BIERNACKI. *Classification simultanée de plusieurs échantillons sous contrainte d'égalité des entropies de partition*, in "Journal de la Société Française de Statistique", 2011.

Invited Conferences

[18] A. AGUILERA, M. ESCABIAS, C. PREDA, G. SAPORTA. *Functional PLS versus functional PCR through simulated data and chemometric applications*, in "4th international Conference of ERCIM WG on Computing and Statistics (ERCIM'11)", 2011.

[19] C. BIERNACKI, V. VANDEWALLE. *Label Switching in Mixtures*, in "AIP Conference Proceedings", AIP, 2011, vol. 1389, p. 398–401 [*DOI :* 10.1063/1.3636747], http://link.aip.org/link/?APC/1389/398/1.

[20] C. PREDA, A. AMARIOAREI. *Approximations for the three-dimensional scan statistics*, in "International Conference on Advances in Probability and Statistics - Theory and Applications", 2011.

[21] C. PREDA. *Functional PLS regression*, in "7th Congress of Romanian Mathematicians", 2011.

International Conferences with Proceedings

[22] C. BOUVEYRON, J. JACQUES. *Model-based Clustering of Time Series in Group-specific Functional Subspaces*, in "12th annual conference of the International Federation of Classification Societies", 2011.

[23] M. GIACOFCI, S. LAMBERT-LACROIX, G. MAROT, F. PICARD. *Wavelet based clustering for mixed-effects functional models*, in "International Biometric Society Channel Network conference", 2011.

[24] A. LOURME, C. BIERNACKI. *Simultaneous t-Model-Based Clustering Applied to Company Bankrupt Prediction*, in "8th Scientific Meeting of the CLAssification and Data Analysis Group of the Italian Statistical Society", 2011.

[25] C. PREDA, J. SCHILTZ. *Functional PLS regression with functional response*, in "Applied Stochastic Models and Data Analysis (ASMDA 2011)", 2011, p. 1126-1133.

National Conferences with Proceeding

[26] C. BIERNACKI, V. VANDEWALLE. *Label switching dans les mélanges*, in "43e Journées de Statistique", 2011.

Conferences without Proceedings

[27] C. BIERNACKI, G. CELEUX, G. GOVAERT, F. LANGROGNET. *Classification des données quantitatives de grande dimension dans l'environnement logiciel MIXMOD*, in "43e Journées de Statistique", 2011.

[28] M.-A. DILLIES, G. MAROT. *RNA-seq Data Analysis: Lost in Normalization?*, in "JOBIM – Journées Ouvertes Biologie Informatique Mathématiques", 2011, with members of the Statomique Consortium.

[29] J. HAMON, C. DHAENENS, J. JACQUES, G. EVEN. *Combining combinatorial optimization and statistics to mine high-throughput genotyping data*, in "JOBIM - Journées Ouvertes Biologie Informatique Mathématiques", Paris, France, June 2011, http://hal.inria.fr/hal-00639533/en.

[30] J. JACQUES. *Functional PLS regression*, in "Astrostatistique en France", 2011.

[31] F. PIERRE, A. REBOUL, B. GRENIER-BOLEY, G. MAROT, M. GUEDJ, R. BLERVAQUE, D. HOT, C. PICHON, H. TOUZET, E. PRADEL, F. SEBBANE. *Toward the identification and characterization of the Yersinia pestis RNome produced in vivo*, in "ASM (American Society for Microbiology) Conference on Regulating with RNA in bacteria", 2011.

[32] V. VANDEWALLE, C. BIERNACKI. *Label Switching in Mixtures*, in "Working Group on Model-Based Clustering Summer Session", 2011.

[33] L. YENGO, J. JACQUES, C. BIERNACKI. *A Block Regression approach for Simultaneous Variables Clustering and Selection: Application to Genetic Data*, in "JOBIM - Journées Ouvertes Biologie Informatique Mathématiques", Paris, France, 2011.

Research Reports

[34] C. BIERNACKI, G. CASTELLAN. *A Data-Driven Bound on Variances for Avoiding Degeneracy in Univariate Gaussian Mixtures*, Pub. IRMA Lille, 2011, n^{o} 71-IV.

Other Publications

[35] A. CÉLISSE, J.-J. DAUDIN, L. PIERRE. *Consistency of maximum likelihood and variational estimators in stochastic block model*, 2011, http://arxiv.org/abs/1105.3288.

[36] J. JACQUES, C. PREDA. *Model-based clustering of functional data*, 2011, http://hal.inria.fr/hal-00628247/en.

[37] M. MARBAC-LOURDELLE. *Modèle de mélange pour variables qualitatives reflétant la corrélation entre variables et application à la classification non supervisée*, Université Lille 1, 2011.

[38] G. MAROT. *Modélisation statistique pour l'analyse de données de puces à ADN*, 2011, GEPV laboratory Lille 1 (Génétique et Evolution des Populations Végétales).

[39] G. MAROT. *Présentation de Bioconductor et de son utilisation sur les puces à ADN*, 2011, séminaire du réseau régional d'ingénieurs en bioinformatique de Lille.

[40] L. YENGO, J. JACQUES, C. BIERNACKI. *A block regression approach for simultaneous clustering and variables selection: application to genetic data*, 2011, Seminar Statistics for Systems Biology (SSB).