IN PARTNERSHIP WITH:
CNRS

Université de Lille

Activity Report 2019

Project-Team MODAL

MOdel for Data Analysis and Learning

IN COLLABORATION WITH: Laboratoire Paul Painlevé (LPP)

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, machine learning and statistical methods

Table of contents

1. Team, Visitors, External Collaborators **1**
2. Overall Objectives **2**
2.1. Context 2
2.2. Goals 3
3. Research Program **3**
3.1. Research axis 1: Unsupervised learning 3
3.2. Research axis 2: Performance assessment 3
3.3. Research axis 3: Functional data 4
3.4. Research axis 4: Applications motivating research 4
4. Application Domains **4**
4.1. Economic world 4
4.2. Biology 4
5. Highlights of the Year **4**
6. New Software and Platforms **5**
6.1. MixtComp 5
6.2. BlockCluster 5
6.3. CloHe 6
6.4. PACBayesianNMF 6
6.5. pycobra 6
6.6. STK++ 6
6.7. rtkore 7
6.8. MixAll 7
6.9. simerge 7
6.10. MixtComp.V4 7
6.11. MASSICCC 8
6.12. Platforms 8
7. New Results **8**
7.1. Axis 1: Data Units Selection in Statistics 8
7.2. Axis 1: Model-Based Co-clustering for Ordinal Data of different dimensions 9
7.3. Axis 1: Model-based co-clustering for mixed type data 9
7.4. Axis 1: Relaxing the Identically Distributed Assumption in Gaussian Co-Clustering for High Dimensional Data 9
7.5. Axis 1: Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering 10
7.6. Axis 1: Co-clustering: A versatile way to perform clustering 10
7.7. Axis 1: Dealing with missing data in model-based clustering through a MNAR model 10
7.8. Axis 1: Organized Co-Clustering for textual data synthesis 11
7.9. Axis 1: Model-Based Co-clustering with Co-variables 11
7.10. Axis 1: Linking canonical and spectral clustering 11
7.11. Axis 1: Predictive clustering 12
7.12. Axis 1: Ranking and synchronization from pairwise measurements via SVD 12
7.13. Axis 1: SPONGE: A generalized eigenproblem for clustering signed networks 12
7.14. Axis 2: Multi-kernel unmixing and super-resolution using the Modified Matrix Pencil method 13
7.15. Axis 2: Provably robust estimation of modulo 1 samples of a smooth function with applications to phase unwrapping 13
7.16. Axis 2: Learning general sparse additive models from point queries in high dimensions 13
7.17. Axis 2: Sparse non-negative super-resolution - simplified and stabilized 14
7.18. Axis 2: Pseudo-Bayesian learning with kernel Fourier transform as prior 14
7.19. Axis 2: PAC-Bayesian binary activated deep neural networks 15
7.20. Axis 2: Improved PAC-Bayesian Bounds for Linear Regression 15

7.21. Axis 2: Multiview Boosting by controlling the diversity and the accuracy of view-specific voters 15
7.22. Axis 2: PAC-Bayes and Domain Adaptation 15
7.23. Axis 2: Interpreting Neural Networks as Majority Votes through the PAC-Bayesian Theory 16
7.24. Axis 2: Still no free lunches: the price to pay for tighter PAC-Bayes bounds 16
7.25. Axis 2: PAC-Bayesian Contrastive Unsupervised Representation Learning 16
7.26. Axis 2: Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly 17
7.27. Axis 2: PAC-Bayes Un-Expected Bernstein Inequality 17
7.28. Axis 2: Attributing and Referencing (Research) Software: Best Practices and Outlook from Inria 17
7.29. Axis 2: Revisiting clustering as matrix factorisation on the Stiefel manifold 18
7.30. Axis 2: A Primer on PAC-Bayesian Learning 18
7.31. Axis 2: Perturbed Model Validation: A New Framework to Validate Model Relevance 18
7.32. Axis 2: Decentralized learning with budgeted network load using Gaussian copulas and classifier ensembles 19
7.33. Axis 2: Online k-means Clustering 19
7.34. Axis 2: Non-linear aggregation of filters to improve image denoising 19
7.35. Axis 2: Multiple change-points detection with reproducing kernels 20
7.36. Axis 2: Analysis of early stopping rules based on discrepancy principle 20
7.37. Axis 3: Short-term air temperature forecasting using Nonparametric Functional Data Analysis and SARMA models 20
7.38. Axis 3: Mathematical Modeling and Study of Random or Deterministic Phenomena 20
7.39. Axis 3: Categorical functional data analysis 21
7.40. Axis 4: Proteomic signature of early death in heart failure patients 21
7.41. Axis 4: Statistical analysis of high-throughput proteomic data 21
7.42. Axis 4: Linking different kinds of Omics data through a model-based clustering approach 22
7.43. Axis 4: Real-time Audio Sources Classification 22
7.44. Axis 4: Matching of descriptors evolving over time 22
7.45. Axis 4: Supervised multivariate discretization and levels merging for logistic regression 23
7.46. Axis 4: MASSICCC Platform for SaaS Software Availability 23
7.47. Axis 4: Domain adaptation from a pre-trained source model 23
7.48. Axis 4: Reject Inference Methods in Credit Scoring: a rational review 24
7.49. Other: Projection Under Pairwise Control 24
8. Bilateral Contracts and Grants with Industry **24**
8.1. Bilateral Contracts with Industry 24
8.2. Bilateral Grants with Industry 24
9. Partnerships and Cooperations **25**
9.1. Regional Initiatives 25
9.2. National Initiatives 25
9.2.1. Programme of Investments for the Future (PIA) 25
9.2.2. RHU PreciNASH 25
9.2.3. CNRS PEPS Blanc – BayesRealForRNN project 25
9.2.4. CNRS AMIES PEPS 2 - DiagChange project 26
9.2.5. AMIES PEPS 1 - CADIS2 26
9.2.6. AMIES PEPS 2 - MadiPa 26
9.2.7. ANR 26
9.2.7.1. ANR APRIORI 26
9.2.7.2. ANR BEAGLE 27
9.2.7.3. ANR SMILE 27
9.2.7.4. ANR TheraSCUD2022 27
9.2.8. Working groups 27

9.2.9. Other initiatives 27
9.3. European Initiatives 28
9.3.1. FP7 & H2020 Projects 28
9.3.2. Collaborations with Major European Organizations 28
9.4. International Initiatives 28
9.4.1. Inria International Labs 28
9.4.2. Inria International Partners 29
9.5. International Research Visitors 29
9.5.1. Visits of International Scientists 29
9.5.2. Visits to International Teams 29
9.5.2.1. Sabbatical programme 29
9.5.2.2. Research Stays Abroad 30
10. Dissemination **30**
10.1. Promoting Scientific Activities 30
10.1.1. Scientific Events: Organisation 30
10.1.2. Scientific Events: Selection 30
10.1.2.1. Member of the Conference Program Committees 30
10.1.2.2. Reviewer 30
10.1.3. Journal 31
10.1.4. Invited Talks 31
10.1.5. Leadership within the Scientific Community 32
10.1.6. Scientific Expertise 32
10.1.7. Research Administration 32
10.2. Teaching - Supervision - Juries 32
10.2.1. Teaching 32
10.2.2. Supervision 33
10.2.2.1. PhD defense: 33
10.2.2.2. PhD in progress: 33
10.2.3. Juries 34
10.3. Popularization 34
10.3.1. Internal or external Inria responsibilities 34
10.3.2. Interventions 34
11. Bibliography **34**

Project-Team MODAL

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

Keywords:

Computer Science and Digital Science:

A3.1.4. - Uncertain data
A3.2.3. - Inference
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.5. - Bayesian methods
A3.4.7. - Kernel methods
A5.2. - Data visualization
A6.2.3. - Probabilistic methods
A6.2.4. - Statistical methods
A6.3.3. - Data processing
A9.2. - Machine learning

Other Research Topics and Application Domains:

B2.2.3. - Cancer
B9.5.6. - Data science
B9.6.3. - Economy, Finance
B9.6.5. - Sociology

1. Team, Visitors, External Collaborators

Research Scientists

Christophe Biernacki [Team leader, Inria, Senior Researcher, HDR]
Pascal Germain [Inria, Researcher, until Oct 2019]
Benjamin Guedj [Inria, Researcher]
Hemant Tyagi [Inria, Researcher]

Faculty Members

Alain Celisse [Université de Lille, Associate Professor, HDR]
Sophie Dabo-Niang [Université de Lille, Professor]
Philippe Heinrich [Université de Lille, Associate Professor]
Serge Iovleff [Université de Lille, Associate Professor]
Guillemette Marot [Université de Lille, Associate Professor]
Cristian Preda [Université de Lille, Professor, HDR]
Vincent Vandewalle [Université de Lille, Associate Professor]

Post-Doctoral Fellows

Florent Dewez [Inria, from Feb 2019]
Fabien Laporte [Ecole polytechnique, until Jul 2019]
Vera Shalaeva [Inria]

PhD Students

Filippo Antonazzo [Inria, from Oct 2019]
Yaroslav Averyanov [Inria]
Maxime Baelde [A-Volute]
Anne-Lise Bedenel [meilleureassurance.com]
Felix Biggs [University College London, UK, from Sep 2019]
Adrien Ehrhardt [CACF, until Mar 2019]
Arthur Leroy [Université de Paris]
Le Li [iAdvize, Université d'Angers]
Antoine Vendeville [University College London, UK, from Sep 2019]
Luxin Zhang [Worldline, from Feb 2019]
Wilfried Heyse [Inria, PhD Student, from Sep 2019]

Technical staff

Iheb Eladib [Inria, Engineer, from Sep 2019]
Olivier Gauriau [Inria, Engineer, from Sep 2019]
Vincent Kubicki [Inria, Engineer, until Jul 2019]
Arthur Talpaert [Inria, Engineer, from Oct 2019]
Jean Francois Bouin [Inria, Engineer, from Apr 2019 until Jun 2019]
Margot Correard [Inria, Engineer, from Apr 2019 until Jun 2019]
Quentin Grimonprez [Inria, Engineer, from Oct 2019]

Interns and Apprentices

Nicolas Bernard [Inria, from Aug 2019 until Sep 2019]
Iheb Eladib [Inria, from Feb 2019 until Jul 2019]
Nathan Forestier [Inria, from Apr 2019 until Jun 2019]
Olivier Gauriau [Inria, from May 2019 until Aug 2019]
Wilfried Heyse [Inria, from Mar 2019 until Aug 2019]
Axel Potier [ADEO, Intern, from Apr 2019 until Sep 2019]
Louis Pujol [Inria, from Apr 2019 until Sep 2019]
Maxime Haddouche [Inria, from Oct 2019]
Tayeb Zarrouk [Inria, from May 2019 until Aug 2019]

Administrative Assistant

Anne Rejl [Inria]

Visiting Scientists

Mihai Cucuringu [University of Oxford, UK, Associate Professor, until Jan 2019]
Abdou Ka Diongue [Gaston Berger University, Senegal, Professor, from Jun 2019 until Jul 2019]
Apoorv Vikram Singh [Independent, Visiting Researcher, from Oct 2019]
Vlad Barbu [Université de Rouen, Associate Professor, from Sep 2019]
Seydou-Nourou Sylla [Independent, Engineer, from May 2019 until Jun 2019]

External Collaborators

Jean Francois Bouin [DiagRAMS, from Jul 2019]
Margot Correard [DiagRAMS, from Jul 2019]

2. Overall Objectives

2.1. Context

In several respects, modern society has strengthened the need for statistical analysis, even if other related names are sometimes preferably used depending on methods, communities and applications, as data analysis, machine learning or artificial intelligence. The genesis comes from the easier availability of data thanks to technological breakthroughs (storage, transfer, computing), and are now so widespread that they are no longer limited to large human organizations. The more or less conscious goal of such data availability is

the expectation to improving the quality of "since the dawn of time" statistical stories which are namely discovering new knowledge or doing better predictions. These both central tasks can be referred respectively as unsupervised learning or supervised learning, even if it is not limited to them or other names exist depending on communities. Somewhere, it pursues the following hope: "more data for better quality and more numerous results".

However, today's data are increasingly complex. They gather mixed type features (for instance continuous data mixed with categorical data), missing or partially missing items (like intervals) and numerous variables (high dimensional situation). As a consequence, the target "better quality and more numerous results" of the previous adage (both words are important: "better quality" and also "more numerous") could not be reached through a somewhat "handwork" way, but should inevitably rely on some theoretical formalization and guarantee. Indeed, data can be so numerous and so complex (data can live in quite abstract spaces) that the "empirical" statistician is quickly outdated. However, data being subject by nature to randomness, the probabilistic framework is a very sensible theoretical environment to serve as a general guide for modern statistical analysis.

2.2. Goals

Modal is a project-team working on today's complex data sets (mixed data, missing data, high-dimensional data), for classical statistical targets (unsupervised learning, supervised learning, regression,...) with approaches relying on the probabilistic framework. This latter can be tackled through both model-based methods (as mixture models for a generic tool) and model-free methods (as probabilistic bounds on empirical quantities). Furthermore, Modal is connected to the real world by applications, typically with biological ones (some members have this skill) but many other ones are also considered since the application coverage of the Modal methodology is very large. It is also important to note that, in return, applications are often real opportunities for initiating academic questioning for the statistician (case of the Bilille platform and some bilateral contracts of the team).

From the academic communities point of view, Modal can be seen as belonging simultaneously to both the statistical learning and machine learning ones, as attested by its publications. Somewhere it is the opportunity to make a bridge between these two stochastic communities around a common but large probabilistic framework.

3. Research Program

3.1. Research axis 1: Unsupervised learning

Scientific locks related to unsupervised learning are numerous, concerning the clustering outcome validity, the ability to manage different kinds of data, the missing data questioning, the dimensionality of the data set,... Many of them are addressed by the team, leading to publication achievements, often with a specific package delivery (sometimes upgraded as a software or even as a platform grouping several software). Because of the variety of the scope, it involves nearly all the permanent team members, often with PhD students and some engineers. The related works are always embedded inside a probabilistic framework, typically model-based approaches but also model-free ones like PAC-Bayes (PAC stands for Probably Approximately Correct), because such a mathematical environment offers both a well-posed problem and a rigorous answer.

3.2. Research axis 2: Performance assessment

One main concern of the Modal team is to provide theoretical justifications on the procedures which are designed. Such guarantees are important to avoid misleading conclusions resulting from any unsuitable use. For example, one ingredient in proving these guarantees is the use of the PAC framework, leading to finite-sample concentration inequalities. More precisely, contributions to PAC learning rely on the classical empirical process theory and the PAC-Bayesian theory. The Modal team exploits such non-asymptotic tools to analyze

the performance of iterative algorithms (such as gradient descent), cross-validation estimators, online change-point detection procedures, ranking algorithms, matrix factorization techniques and clustering methods, for instance. The team also develops some expertise on the formal dynamic study of algorithms related to mixture models (important models used in the previous unsupervised setting), like degeneracy for EM algorithm or also label switching for Gibbs algorithm.

3.3. Research axis 3: Functional data

Mainly due to technological advances, functional data are more and more widespread in many application domains. Functional data analysis (FDA) is concerned with the modeling of data, such as curves, shapes, images or a more complex mathematical object, though as smooth realizations of a stochastic process (an infinite dimensional data object valued in a space of eventually infinite dimension; space of squared integrable functions,...). Time series are an emblematic example even if it should not be limited to them (spectral data, spatial data,...). Basically, FDA considers that data correspond to realizations of stochastic processes, usually assumed to be in a metric, semi-metric, Hilbert or Banach space. One may consider, functional independent or dependent (in time or space) data objects of different types (qualitative, quantitative, ordinal, multivariate, time-dependent, spatial-dependent,...). The last decade saw a dynamic literature on parametric or non-parametric FDA approaches for different types of data and applications to various domains, such as principal component analysis, clustering, regression and prediction.

3.4. Research axis 4: Applications motivating research

The fourth axis consists in translating real application issues into statistical problems raising new (academic) challenges for models developed in Modal team. Cifre Phds in industry and interdisciplinary projects with research teams in Health and Biology are at the core of this objective. The main originality of this objective lies in the use of statistics with complex data, including in particular ultra-high dimension problems. We focus on real applications which cannot be solved by classical data analysis.

4. Application Domains

4.1. Economic world

The Modal team applies it research to the economic world through CIFRE Phd supervision such as CACF (credit scoring), A-Volute (expert in 3D sound), Meilleur Taux (insurance comparator), Worldline. It also has several contracts with companies such as COLAS, Nokia-Apsys/Airbus.

4.2. Biology

The second main application domain of the team is the biology. Members of the team are involved in the supervision and scientific animation of the bilille platform, the bioinformatics and bioanalysis platform and OncoLille project of Lille.

5. Highlights of the Year

5.1. Highlights of the Year

- Benjamin Guedj gave (with John Shawe-Taylor) a plenary tutorial of 2 hours for opening the ICML 2019 (Longbeach, California, USA – June 2019).
- Official creation in July 2019 of a startup DiagRAMS using MODAL's technology (MixtComp software) for predictive maintenance.

- Benjamin Guedj has received two best reviewer awards (top 5% of reviewers) for ICML 2019 and NeurIPS 2019, the flagship conferences in machine learning. Pascal Germain received the best reviewer award (top 5% of reviewers) for NeurIPS 2019.

5.1.1. More relevant results in 2019.

While Section 7 contains a complete list of results for 2019, the important results which were published in peer-reviewed international conferences/journals are described in Sections 7.2, 7.3, 7.13, 7.16, 7.17, 7.18, 7.19, 7.21, 7.22, 7.27, 7.28, 7.32, 7.37, 7.43 and 7.47.

6. New Software and Platforms

6.1. MixtComp

Mixture Computation

KEYWORDS: Clustering - Statistics - Missing data

FUNCTIONAL DESCRIPTION: MixtComp (Mixture Computation) is a model-based clustering package for mixed data originating from the Modal team (Inria Lille). It has been engineered around the idea of easy and quick integration of all new univariate models, under the conditional independence assumption. New models will eventually be available from researches, carried out by the Modal team or by other teams. Currently, central architecture of MixtComp is built and functionality has been field-tested through industry partnerships. Three basic models (Gaussian, multinomial, Poisson) are implemented, as well as two advanced models (Ordinal and Rank). MixtComp has the ability to natively manage missing data (completely or by interval). MixtComp is used as an R package, but its internals are coded in C++ using state of the art libraries for faster computation.

- Participants: Christophe Biernacki, Etienne Goffinet, Matthieu Marbac-Lourdelle, Quentin Grimonprez, Serge Iovleff and Vincent Kubicki
- Contact: Christophe Biernacki
- URL: https://cran.r-project.org/web/packages/RMixtComp/index.html

6.2. BlockCluster

Block Clustering

KEYWORDS: Statistic analysis - Clustering package

SCIENTIFIC DESCRIPTION: Simultaneous clustering of rows and columns, usually designated by biclustering, co-clustering or block clustering, is an important technique in two way data analysis. It consists of estimating a mixture model which takes into account the block clustering problem on both the individual and variables sets. The blockcluster package provides a bridge between the C++ core library and the R statistical computing environment. This package allows to co-cluster binary, contingency, continuous and categorical data-sets. It also provides utility functions to visualize the results. This package may be useful for various applications in fields of Data mining, Information retrieval, Biology, computer vision and many more.

FUNCTIONAL DESCRIPTION: BlockCluster is an R package for co-clustering of binary, contingency and continuous data based on mixture models.

RELEASE FUNCTIONAL DESCRIPTION: Initialization strategy enhanced

- Participants: Christophe Biernacki, Gilles Celeux, Parmeet Bhatia, Serge Iovleff, Vincent Brault and Vincent Kubicki
- Partner: Université de Technologie de Compiègne
- Contact: Serge Iovleff
- URL: http://cran.r-project.org/web/packages/blockcluster/index.html

6.3. CloHe

Clustering of Mixed data

KEYWORDS: Classification - Clustering - Missing data

FUNCTIONAL DESCRIPTION: Software of classification for mixed data with missing values with application to multispectral satellite image time-series

- Partners: CNRS - INRA
- Contact: Serge Iovleff
- URL: https://modal.lille.inria.fr/CloHe/

6.4. PACBayesianNMF

KEYWORDS: Statistics - Machine learning

FUNCTIONAL DESCRIPTION: Implementing NMF with a PAC-Bayesian approach relying upon block gradient descent

- Participants: Benjamin Guedj and Astha Gupta
- Contact: Benjamin Guedj
- URL: https://github.com/astha736/PACbayesianNMF

6.5. pycobra

KEYWORDS: Statistics - Data visualization - Machine learning

SCIENTIFIC DESCRIPTION: pycobra is a python library for ensemble learning, which serves as a toolkit for regression, classification, and visualisation. It is scikit-learn compatible and fits into the existing scikit-learn ecosystem.

pycobra offers a python implementation of the COBRA algorithm introduced by Biau et al. (2016) for regression.

Another algorithm implemented is the EWA (Exponentially Weighted Aggregate) aggregation technique (among several other references, you can check the paper by Dalalyan and Tsybakov (2007).

Apart from these two regression aggregation algorithms, pycobra implements a version of COBRA for classification. This procedure has been introduced by Mojirsheibani (1999).

pycobra also offers various visualisation and diagnostic methods built on top of matplotlib which lets the user analyse and compare different regression machines with COBRA. The Visualisation class also lets you use some of the tools (such as Voronoi Tesselations) on other visualisation problems, such as clustering.

- Participants: Bhargav Srinivasa Desikan and Benjamin Guedj
- Contact: Benjamin Guedj
- Publication: Pycobra: A Python Toolbox for Ensemble Learning and Visualisation
- URL: https://github.com/bhargavvader/pycobra

6.6. STK++

Statistical ToolKit

KEYWORDS: Statistics - Linear algebra - Framework - Learning - Statistical learning

FUNCTIONAL DESCRIPTION: STK++ (Statistical ToolKit in C++) is a versatile, fast, reliable and elegant collection of C++ classes for statistics, clustering, linear algebra, arrays (with an API Eigen-like), regression, dimension reduction, etc. The library is interfaced with lapack for many linear algebra usual methods. Some functionalities provided by the library are available in the R environment using rtkpp and rtkore.

STK++ is suitable for projects ranging from small one-off projects to complete data mining application suites.

- Participant: Serge Iovleff
- Contact: Serge Iovleff
- URL: http://www.stkpp.org

6.7. rtkore

STK++ core library integration to R using Rcpp

KEYWORDS: C++ - Data mining - Clustering - Statistics - Regression

FUNCTIONAL DESCRIPTION: STK++ (http://www.stkpp.org) is a collection of C++ classes for statistics, clustering, linear algebra, arrays (with an Eigen-like API), regression, dimension reduction, etc. The integration of the library to R is using Rcpp. The rtkore package includes the header files from the STK++ core library. All files contain only templated classes or inlined functions. STK++ is licensed under the GNU LGPL version 2 or later. rtkore (the stkpp integration into R) is licensed under the GNU GPL version 2 or later. See file LICENSE.note for details.

- Participant: Serge Iovleff
- Contact: Serge Iovleff
- URL: https://cran.r-project.org/web/packages/rtkore/index.html

6.8. MixAll

Clustering using Mixture Models

KEYWORDS: Clustering - Clustering package - Generative Models

FUNCTIONAL DESCRIPTION: MixAll is a model-based clustering package for modelling mixed data sets. It has been engineered around the idea of easy and quick integration of any kind of mixture models for any kind of data, under the conditional independence assumption. Currently five models (Gaussian mixtures, categorical mixtures, Poisson mixtures, Gamma mixtures and kernel mixtures) are implemented. MixAll has the ability to natively manage completely missing values when assumed as random. MixAll is used as an R package, but its internals are coded in C++ as part of the STK++ library (www.stkpp.org) for faster computation.

RELEASE FUNCTIONAL DESCRIPTION: clusterPredict allow to predcit membership for a new data set using a previously estimated model

- Participant: Serge Iovleff
- Partner: Université Lille 1
- Contact: Serge Iovleff
- URL: https://cran.r-project.org/web/packages/MixAll/

6.9. simerge

Statistical Inference for the Management of Extrem Risks, Genetics and Global epidemiology

KEYWORD: Biclustering

FUNCTIONAL DESCRIPTION: Allows to perform Co-Clustering on binary (Bernoulli) and counting variables (Poisson) using co-variables.

- Partner: Inria
- Contact: Serge Iovleff

6.10. MixtComp.V4

KEYWORDS: Clustering - Statistics - Missing data - Mixed data

FUNCTIONAL DESCRIPTION: MixtComp (Mixture Computation) is a model-based clustering package for mixed data originating from the Modal team (Inria Lille). It has been engineered around the idea of easy and quick integration of all new univariate models, under the conditional independence assumption. New models will eventually be available from researches, carried out by the Modal team or by other teams. Currently, central architecture of MixtComp is built and functionality has been field-tested through industry partnerships. Five basic models (Gaussian, Multinomial, Poisson, Weibull, NegativeBinomial) are implemented, as well as two advanced models (Functional and Rank). MixtComp has the ability to natively manage missing data (completely or by interval). MixtComp is used as an R package, but its internals are coded in C++ using state of the art libraries for faster computation.

RELEASE FUNCTIONAL DESCRIPTION: - New I/O system - Replacement of regex library - Improvement of initialization - Criteria for stopping the algorithm - Added management of partially missing data for several models - User documentation - Adding user features in R

- Participants: Christophe Biernacki, Vincent Kubicki, Matthieu Marbac-Lourdelle, Serge Iovleff, Quentin Grimonprez and Etienne Goffinet
- Partners: Université de Lille - CNRS
- Contact: Christophe Biernacki

6.11. MASSICCC

Massive Clustering with Cloud Computing

KEYWORDS: Statistic analysis - Big data - Machine learning - Web Application

SCIENTIFIC DESCRIPTION: The web application let users use several software packages developed by Inria directly in a web browser. Mixmod is a classification library for continuous and categorical data. MixtComp allows for missing data and a larger choice of data types. BlockCluster is a library for co-clustering of data. When using the web application, the user can first upload a data set, then configure a job using one of the libraries mentioned and start the execution of the job on a cluster. The results are then displayed directly in the browser allowing for rapid understanding and interactive visualisation.

FUNCTIONAL DESCRIPTION: The MASSICCC web application offers a simple and dynamic interface for analysing heterogeneous data with a web browser. Various software packages for statistical analysis are available (Mixmod, MixtComp, BlockCluster) which allow for supervised and supervised classification of large data sets.

- Contact: Christophe Biernacki
- URL: https://massiccc.lille.inria.fr

6.12. Platforms

6.12.1. MASSICCC Platform

MASSICCC is a demonstration platform giving access through a SaaS (service as a software) concept to data analysis libraries developed at Inria. It allows obtaining results either directly through a website specific display (specific and interactive visual outputs) or through an R data object download. It started in October 2015 for two years and is common to the Modal team (Inria Lille) and the Select team (Inria Saclay). In 2016, two packages have been integrated: Mixmod and MixtComp (see the specific section about MixtComp). In 2017, the BlockCluster package has been integrated and also a particular attention to provide meaningful graphical outputs (for Mixmod, MixtComp and BlockCluster) directly in the web platform itself has led to some specific developments. In 2019, a new version of the MixtComp software has been developed.

7. New Results

7.1. Axis 1: Data Units Selection in Statistics

Participant: Christophe Biernacki.

Usually, the data unit definition is fixed by the practitioner but it can happen that he/she hesitates between several data unit options. In this context, it is highlighted that it is possible to embed data unit selection into a classical model selection principle. The problem is introduced in a regression context before to focus on the model-based clustering and co-clustering context, for data of different kinds (continuous, count, categorical). This work was published in an international journal in 2018 and leads to a keynote as an invited speaker to the 12th Scientific Meeting Classification and Data Analysis Group Cassino (CLADAG 2019) in Italy [41].

It is a joint work with Alexandre Lourme from University of Bordeaux.

7.2. Axis 1: Model-Based Co-clustering for Ordinal Data of different dimensions

Participant: Christophe Biernacki.

This work has been motivated by a psychological survey on women affected by a breast tumor. Patients replied at different moments of their treatment to questionnaires with answers on ordinal scale. The questions relate to aspects of their life called dimensions. To assist the psychologists in analyzing the results, it is useful to emphasize a structure in the dataset. The clustering method achieves that by creating groups of individuals that are depicted by a representative of the group. From a psychological position, it is also useful to observe how questions may be grouped. This is why a clustering should also be performed on the features, which is called a co-clustering problem. However, gathering questions that are not related to the same dimension does not make sense from a psychologist stance. Therefore, the present work corresponds to perform a constrained co-clustering method aiming to prevent questions from different dimensions from getting assembled in a same column-cluster. In addition, evolution of co-clusters along time has been investigated. The method relies on a constrained Latent Block Model embedding a probability distribution for ordinal data. Parameter estimation relies on a Stochastic EM-algorithm associated to a Gibbs sampler, and the ICL-BIC criterion is used for selecting the numbers of co-clusters. The resulting work is now accepted in an international journal [28]. The related R package ordinalClust has been also written and has led to a specific preprint [73] now submitted to an international journal.

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2, and Florence Cousson-Gélie from University Paul Valéry Montpellier 3.

7.3. Axis 1: Model-based co-clustering for mixed type data

Participant: Christophe Biernacki.

Over decades, a lot of studies have shown the importance of clustering to emphasize groups of observations. More recently, due to the emergence of high-dimensional datasets with a huge number of features, co-clustering techniques have emerged and proposed several methods for simultaneously producing groups of observations and features. By synthesizing the dataset in blocks (the crossing of a row-cluster and a column-cluster), this technique can sometimes summarize better the data and its inherent structure. The Latent Block Model (LBM) is a well-known method for performing a co-clustering. However, recently, contexts with features of different types (here called mixed type datasets) are becoming more common. Unfortunately, the LBM is not directly applicable on this kind of dataset. The present work extends the usual LBM to the so-called Multiple Latent Block Model (MLBM) which is able to handle mixed type datasets. The inference is done through a Stochastic EM-algorithm embedding a Gibbs sampler and model selection criterion is defined to choose the number of row and column clusters. This method was successfully used on simulated and real datasets. This work is now accepted in an international journal [29].

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2.

7.4. Axis 1: Relaxing the Identically Distributed Assumption in Gaussian Co-Clustering for High Dimensional Data

Participant: Christophe Biernacki.

A co-clustering model for continuous data that relaxes the identically distributed assumption within blocks of traditional co-clustering is presented. The proposed model, although allowing more flexibility, still maintains the very high degree of parsimony achieved by traditional co-clustering. A stochastic EM algorithm along with a Gibbs sampler is used for parameter estimation and an ICL criterion is used for model selection. Simulated and real datasets are used for illustration and comparison with traditional co-clustering. This work has been submitted to an international journal [63].

This is a joint work with Michael Gallaugher (PhD student) and Paul McNicholas, both from McMaster University (Canada). Michael Gallaugher visited Modal for three months in 2018.

7.5. Axis 1: Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

A generic method is introduced to visualize in a Gaussian-like way, and onto R^2, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have a thorough overview of the potentially complex clustering result. An entropic measure allows us to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis. This work is now in minor revision for an international journal [54]. It has also led to an invited talk to an international conference [42], and several other invitations (the workshop "Advances in data science for big and complex data" at Université Paris-Dauphine in January and the seminary of the Probability and Statistics team of the University Nice Sophia-Antipolis in November).

This is a joint work with Matthieu Marbac from ENSAI.

7.6. Axis 1: Co-clustering: A versatile way to perform clustering

Participant: Christophe Biernacki.

Standard model-based clustering is known to be very efficient for low dimensional data sets, but it fails for properly addressing high dimension (HD) ones, where it suffers from both statistical and computational drawbacks. In order to counterbalance this curse of dimensionality, some proposals have been made to take into account redundancy and features utility, but related models are not suitable for too many variables. We advocate that the latent bloc model, a probabilistic model for co-clustering, is of particular interest to perform HD clustering of individuals even if it is not its primary function. We illustrate in an empirical manner the trade-off bias-variance of the co-clustering strategy in scenarii involving HD fundaments (correlated variables, irrelevant variables) and show the ability of co-clustering to outperform simple mixture row-clustering. An early version of this work has been presented to an national conference with international audience [46].

We also co-organized a special session to an international conference [45] to discuss the potential links between deterministic methods for co-clustering (based on a metric and computer science procedure) or probabilistic methods for co-clustering (mainly based on mixture models). It was the opportunity to gather related communities which are often distinct.

All are joint works with Christine Keribin from Université Paris-Sud.

7.7. Axis 1: Dealing with missing data in model-based clustering through a MNAR model

Participants: Christophe Biernacki, Fabien Laporte.

Since the 90s, model-based clustering is largely used to classify data. Nowadays, with the increase of available data, missing values are more frequent. Traditional ways to deal with them consist in obtaining a filled data set, either by discarding missing values or by imputing them. In the first case, some information is lost; in the second case, the final clustering purpose is not taken into account through the imputation step. Thus, both solutions risk to blur the clustering estimation result. Alternatively, we defend the need to embed the missingness mechanism directly within the clustering modeling step. There exists three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In all situations logistic regression is proposed as a natural and flexible candidate model. In particular, its flexibility property allows us to design some meaningful parsimonious variants, as dependency on missing values or dependency on the cluster label. In this unified context, standard model selection criteria can be used to select between such different missing data mechanisms, simultaneously with the number of clusters. Practical interest of our proposal is illustrated on data derived from medical studies suffering from many missing data. This work has been presented as an invited speaker to an international conference [31]. It has also been presented at a national conference with international audience [47] and as a poster to the international Working Group on Model-Based Clustering [69]. Currently, a preprint is being finalized for submission to an international journal.

It is a joint work with Gilles Celeux from Inria Saclay and Julie Josse from Ecole Polytechnique.

7.8. Axis 1: Organized Co-Clustering for textual data synthesis

Participant: Christophe Biernacki.

Recently, different studies have demonstrated the interest of co-clustering, which simultaneously produces clusters of lines and columns. The present work introduces a novel co-clustering model for parsimoniously summarizing textual data in documents × terms format. Besides highlighting homogeneous coclusters - as other existing algorithms do - we also distinguish noisy coclusters from significant ones, which is particularly useful for sparse documents × term matrices. Furthermore, our model proposes a structure among the significant coclusters and thus obtains a better interpretability to the user. By forcing a structure through row-clusters and column-clusters, this approach is competitive in terms of documents clustering, and offers user-friendly results. The algorithm derived for the proposed method is a Stochastic EM algorithm embedding a Gibbs sampling step and the Poisson distribution. A paper is currently in revision in an international journal [72].

This is joint work with Margot Selosse (PhD student) and Julien Jacques, both from University of Lyon 2.

7.9. Axis 1: Model-Based Co-clustering with Co-variables

Participant: Serge Iovleff.

This work has been motivated by an epidemiological and genetic survey of malaria disease in Senegal. Data were collected between 1990 and 2008. It is based on a latent block model taking into account the problem of grouping variables and clustering individuals by integrating information given by a set of co-variables. Numerical experiments on simulated data sets and an application on real genetic data highlight the interest of this approach. An article has been submitted to *Journal of Classification* and should incorporate “Major Revisions”.

7.10. Axis 1: Linking canonical and spectral clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

It is a recent work aiming at defining a mathematical bridge between classical model-based clustering and classical spectral clustering. Interest of such a prospect is to be able to compare both methods through the rigorous scheme of model selection paradigm. It is still an ongoing work, with several short working papers.

It is a joint work with Alexandre Lourme from University of Bordeaux.

7.11. Axis 1: Predictive clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

Many data, for instance in biostatistics, contain some sets of variables which permit evaluating unobserved traits of the subjects (e.g., we ask question about how many pizzas, hamburgers, chips... are eaten to know how healthy are the food habits of the subjects). Moreover, we often want to measure the relations between these unobserved traits and some target variables (e.g., obesity). Thus, a two-steps procedure is often used: first, a clustering of the observations is performed on the sets of variables related to the same topic; second, the predictive model is fitted by plugging the estimated partitions as covariates. Generally, the estimated partitions are not exactly equal to the true ones. We investigate the impact of these measurement errors on the estimators of the regression parameters, and we explain when this two-steps procedure is consistent. We also present a specific EM algorithm which simultaneously estimates the parameters of the clustering and predictive models. It is an ongoing work.

It is a joint work with Matthieu Marbac from ENSAI and Mohammed Sedki from University Paris-Sud.

7.12. Axis 1: Ranking and synchronization from pairwise measurements via SVD

Participant: Hemant Tyagi.

Given a measurement graph $G = ([n], E)$ and an unknown signal $r \in R^n$, we investigate algorithms for recovering r from pairwise measurements of the form $r_i - r_j$; $\{i, j\} \in E$. This problem arises in a variety of applications, such as ranking teams in sports data and time synchronization of distributed networks. Framed in the context of ranking, the task is to recover the ranking of n teams (induced by r) given a small subset of noisy pairwise rank offsets. We propose a simple SVD-based algorithmic pipeline for both the problem of time synchronization and ranking. We provide a detailed theoretical analysis in terms of robustness against both sampling sparsity and noise perturbations with outliers, using results from matrix perturbation and random matrix theory. Our theoretical findings are complemented by a detailed set of numerical experiments on both synthetic and real data, showcasing the competitiveness of our proposed algorithms with other state-of-the-art methods.

This is joint work with Alexandre d'Aspremont (CNRS & ENS, Paris) and Mihai Cucuringu (University of Oxford, UK) and is available as a preprint [61].

7.13. Axis 1: SPONGE: A generalized eigenproblem for clustering signed networks

Participant: Hemant Tyagi.

We introduce a principled and theoretically sound spectral method for k-way clustering in signed graphs, where the affinity measure between nodes takes either positive or negative values. Our approach is motivated by social balance theory, where the task of clustering aims to decompose the network into disjoint groups, such that individuals within the same group are connected by as many positive edges as possible, while individuals from different groups are connected by as many negative edges as possible. Our algorithm relies on a generalized eigenproblem formulation inspired by recent work on constrained clustering. We provide theoretical guarantees for our approach in the setting of a signed stochastic block model, by leveraging tools from matrix perturbation theory and random matrix theory. An extensive set of numerical experiments on both synthetic and real data shows that our approach compares favorably with state-of-the-art methods for signed clustering, especially for large number of clusters and sparse measurement graphs.

This is joint work with Mihai Cucuringu (University of Oxford, UK), Peter Davies (University of Warwick, UK) and Aldo Glielmo (Imperial College, London, UK) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It was published in the proceedings of an international conference [32].

7.14. Axis 2: Multi-kernel unmixing and super-resolution using the Modified Matrix Pencil method

Participant: Hemant Tyagi.

Consider L groups of point sources or spike trains, with the l^{th} group represented by $x_l(t)$. For a function $g : R \to R$, let $g_l(t) = g(t/\mu_l)$ denote a point spread function with scale $\mu_l > 0$, and with $\mu_1 < \cdots < \mu_L$. With $y(t) = \sum_{l=1}^{L} (g_l \bar{\star} x_l)(t)$, our goal is to recover the source parameters given samples of y, or given the Fourier samples of y. This problem is a generalization of the usual super-resolution setup wherein $L = 1$; we call this the multi-kernel unmixing super-resolution problem. Assuming access to Fourier samples of y, we derive an algorithm for this problem for estimating the source parameters of each group, along with precise non-asymptotic guarantees. Our approach involves estimating the group parameters sequentially in the order of increasing scale parameters, i.e., from group 1 to L. In particular, the estimation process at stage $1 \leq l \leq L$ involves (i) carefully sampling the tail of the Fourier transform of y, (ii) a *deflation* step wherein we subtract the contribution of the groups processed thus far from the obtained Fourier samples, and (iii) applying Moitra's modified Matrix Pencil method on a deconvolved version of the samples in (ii).

This is joint work with Stephane Chretien (National Physical Laboratory, UK & Alan Turing Institute, London) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It is currently under revision in an international journal and is available as a preprint [56].

7.15. Axis 2: Provably robust estimation of modulo 1 samples of a smooth function with applications to phase unwrapping

Participant: Hemant Tyagi.

Consider an unknown smooth function $f : [0,1]^d \to R$, and assume we are given n noisy mod 1 samples of f, i.e., $y_i = (f(x_i) + \eta_i) \mod 1$, for $x_i \in [0,1]^d$, where η_i denotes the noise. Given the samples $(x_i, y_i)_{i=1}^n$, our goal is to recover smooth, robust estimates of the clean samples $f(x_i) \mod 1$. We formulate a natural approach for solving this problem, which works with angular embeddings of the noisy mod 1 samples over the unit circle, inspired by the angular synchronization framework. This amounts to solving a smoothness regularized least-squares problem – a quadratically constrained quadratic program (QCQP) – where the variables are constrained to lie on the unit circle. Our proposed approach is based on solving its relaxation, which is a *trust-region sub-problem* and hence solvable efficiently. We provide theoretical guarantees demonstrating its robustness to noise for adversarial, as well as random Gaussian and Bernoulli noise models. To the best of our knowledge, these are the first such theoretical results for this problem. We demonstrate the robustness and efficiency of our proposed approach via extensive numerical simulations on synthetic data, along with a simple least-squares based solution for the unwrapping stage, that recovers the original samples of f (up to a global shift). It is shown to perform well at high levels of noise, when taking as input the denoised modulo 1 samples. Finally, we also consider two other approaches for denoising the modulo 1 samples that leverage tools from Riemannian optimization on manifolds, including a Burer-Monteiro approach for a semidefinite programming relaxation of our formulation. For the two-dimensional version of the problem, which has applications in synthetic aperture radar interferometry (InSAR), we are able to solve instances of real-world data with a million sample points in under 10 seconds, on a personal laptop.

This is joint work with Mihai Cucuringu (University of Oxford, UK) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It has been accepted to appear (after minor revision) in an international journal, and is available as a preprint [60].

7.16. Axis 2: Learning general sparse additive models from point queries in high dimensions

Participant: Hemant Tyagi.

We consider the problem of learning a d-variate function f defined on the cube $[-1,1]^d \subset R^d$, where the algorithm is assumed to have black box access to samples of f within this domain. Denote $S_r; r = 1, \cdots, r_0$ to be sets consisting of unknown r-wise interactions amongst the coordinate variables. We then focus on the setting where f has an additive structure, i.e., it can be represented as

$$f = \sum_{j \in \S_1} \phi_j + \sum_{j \in S_2} \phi_j + \cdots + \sum_{j \in S_{r_0}} \phi_j,$$

where each ϕ_j; $j \in S_r$ is at most r-variate for $1 \leq r \leq r_0$. We derive randomized algorithms that query f at carefully constructed set of points, and exactly recover each S_r with high probability. In contrary to the previous work, our analysis does not rely on numerical approximation of derivatives by finite order differences.

This is joint work with Jan Vybiral (Czech Technical University, Prague) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It has now been published in an international journal [30].

7.17. Axis 2: Sparse non-negative super-resolution - simplified and stabilized

Participant: Hemant Tyagi.

The convolution of a discrete measure, $x = \sum_{i=1}^{k} a_i \delta_{t_i}$, with a local window function, $\phi(s-t)$, is a common model for a measurement device whose resolution is substantially lower than that of the objects being observed. Super-resolution concerns localising the point sources with an accuracy beyond the essential support of $\phi(s-t)$, typically from m noisy samples of the convolution output. We consider the setting of x being non-negative and seek to characterise all non-negative measures approximately consistent with the samples. We first show that x is the unique non-negative measure consistent with the samples provided the samples are exact, and $m \geq 2k+1$ samples are available, and $\phi(s-t)$ generates a Chebyshev system. This is independent of how close the sample locations are and *does not rely on any regulariser beyond non-negativity*; as such, it extends and clarifies the work by Schiebinger et al. and De Castro et al., who achieve the same results but require a total variation regulariser, which we show is unnecessary. Moreover, we establish stability results in the setting where the samples are corrupted with noise. The main innovation of these results is that non-negativity alone is sufficient to localise point sources beyond the essential sensor resolution.

This is joint work with Armin Eftekhari (EPFL, Switzerland), Jared Tanner (University of Oxford, UK), Andrew Thompson (National Physical Laboratory, UK), Bogdan Toader (University of Oxford, UK) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It has now been published in an international journal [24].

7.18. Axis 2: Pseudo-Bayesian learning with kernel Fourier transform as prior

Participant: Pascal Germain.

We revisit the kernel random Fourier features (RFF) method through the lens of the PAC-Bayesian theory. While the primary goal of RFF is to approximate a kernel, we look at the Fourier transform as a prior distribution over trigonometric hypotheses. It naturally suggests learning a posterior on these hypotheses. We derive generalization bounds that are optimized by learning a pseudo-posterior obtained from a closed-form expression, and corresponding learning algorithms.

This joint work with Emilie Morvant from Université Jean Monnet de Saint-Etienne (France), and Gaël Letarte from Université Laval (Québec, Canada) has been initiated in 2018 when Gaël Letarte was doing an internship at Inria, and led to a publication in the proceedings of AISTATS 2019 conference [36]. The same work has been prensented as a poster in the "Workshop on Machine Learning with guarantees @ NeurIPS 2019".

An extension of this work, co-authored with Léo Gautheron, Amaury Habrard, Marc Sebban, and Valentina Zantedeschi – all from Université Jean Monnet de Saint-Etienne – has been presented at the national conference CAp 2019 [44]. It is also the topic of a technical report [64].

7.19. Axis 2: PAC-Bayesian binary activated deep neural networks

Participant: Pascal Germain, Benjamin Guedj

We present a comprehensive study of multilayer neural networks with binary activation, relying on the PAC-Bayesian theory. Our contributions are twofold: (i) we develop an end-to-end framework to train a binary activated deep neural network, overcoming the fact that binary activation function is non-differentiable; (ii) we provide nonvacuous PAC-Bayesian generalization bounds for binary activated deep neural networks. Noteworthy, our results are obtained by minimizing the expected loss of an architecture-dependent aggregation of binary activated deep neural networks. The performance of our approach is assessed on a thorough numerical experiment protocol on real-life datasets. This work has been published in the proceedings of NeurIPS 2019 conference [35].

It is a joint work with Gaël Letarte and François Laviolette, from Université Laval (Québec, Canada).

7.20. Axis 2: Improved PAC-Bayesian Bounds for Linear Regression

Participant: Pascal Germain, Vera Shalaeva

We improve the PAC-Bayesian error bound for linear regression provided in the literature. The improvements are two-fold. First, the proposed error bound is tighter, and converges to the generalization loss with a well-chosen temperature parameter. Second, the error bound also holds for training data that are not independently sampled. In particular, the error bound applies to certain time series generated by well-known classes of dynamical models, such as ARX models.

It is a joint work with Mihaly Petreczky and Alireza Fakhrizadeh Esfahani from Université de Lille. It has been accepted for publication as part of the AAAI 2020 conference [38].

7.21. Axis 2: Multiview Boosting by controlling the diversity and the accuracy of view-specific voters

Participant: Pascal Germain

We present a comprehensive study of multilayer neural networks with binary activation, relying on the PAC-Bayesian We propose a boosting based multiview learning algorithm which iteratively learns i) weights over view-specific voters capturing view-specific information; and ii) weights over views by optimizing a PAC-Bayes multiview C-Bound that takes into account the accuracy of view-specific classifiers and the diversity between the views. We derive a generalization bound for this strategy following the PAC-Bayes theory which is a suitable tool to deal with models expressed as weighted combination over a set of voters.

It is a joint work with Emilie Morvant from Université Jean Monnet de Saint-Etienne and with Massih-Reza Amini of Université de Grenoble, and with Anil Goyal affiliated to both institutions. This work has been published in the journal Neurocomputing [26].

7.22. Axis 2: PAC-Bayes and Domain Adaptation

Participant: Pascal Germain

In machine learning, Domain Adaptation (DA) arises when the distribution generating the test (target) data differs from the one generating the learning (source) data. It is well known that DA is a hard task even under strong assumptions, among which the covariate-shift where the source and target distributions diverge only in their marginals, i.e. they have the same labeling function. Another popular approach is to consider a hypothesis class that moves closer the two distributions while implying a low-error for both tasks. This is a VC-dim approach that restricts the complexity of a hypothesis class in order to get good generalization. Instead, we propose a PAC-Bayesian approach that seeks for suitable weights to be given to each hypothesis in order to build a majority vote. We prove a new DA bound in the PAC-Bayesian context. This leads us to design the first DA-PAC-Bayesian algorithm based on the minimization of the proposed bound. Doing so, we

seek for a ρ-weighted majority vote that takes into account a trade-off between three quantities. The first two quantities being, as usual in the PAC-Bayesian approach, (a) the complexity of the majority vote (measured by a Kullback-Leibler divergence) and (b) its empirical risk (measured by the ρ-average errors on the source sample). The third quantity is (c) the capacity of the majority vote to distinguish some structural difference between the source and target samples.

This work has been published in the journal Neurocomputing [25].

It is a joint work with Emilie Morvant and Amaury Habrard from Université Jean Monnet de Saint-Etienne (France), and with François Laviolette from Université Laval (Québec, Canada).

7.23. Axis 2: Interpreting Neural Networks as Majority Votes through the PAC-Bayesian Theory

Participant: Pascal Germain, Paul Viallard

We propose a PAC-Bayesian theoretical study of the two-phase learning procedure of a neural network introduced by Kawaguchi et al. (2017). In this procedure, a network is expressed as a weighted combination of all the paths of the network (from the input layer to the output one), that we reformulate as a PAC-Bayesian majority vote. Starting from this observation, their learning procedure consists in (1) learning "prior" network for fixing some parameters, then (2) learning a "posterior" network by only allowing a modification of the weights over the paths of the prior network. This allows us to derive a PAC-Bayesian generalization bound that involves the empirical individual risks of the paths (known as the Gibbs risk) and the empirical diversity between pairs of paths. Note that similarly to classical PAC-Bayesian bounds, our result involves a KL-divergence term between a "prior" network and the "posterior" network. We show that this term is computable by dynamic programming without assuming any distribution on the network weights.

This early result has been accepted as a poster presentation in the international workshop "Workshop on Machine Learning with guarantees @ NeurIPS 2019" [50].

This is a joint work with researchers from Université Jean Monnet de Saint-Etienne: Amaury Habrard, Emilie Morvant, and Rémi Emonet.

7.24. Axis 2: Still no free lunches: the price to pay for tighter PAC-Bayes bounds

Participant: Benjamin Guedj

"No free lunch" results state the impossibility of obtaining meaningful bounds on the error of a learning algorithm without prior assumptions and modelling. Some models are expensive (strong assumptions, such as as subgaussian tails), others are cheap (simply finite variance). As it is well known, the more you pay, the more you get: in other words, the most expensive models yield the more interesting bounds. Recent advances in robust statistics have investigated procedures to obtain tight bounds while keeping the cost minimal. The present paper explores and exhibits what the limits are for obtaining tight PAC-Bayes bounds in a robust setting for cheap models, addressing the question: is PAC-Bayes good value for money?

Joint work with Louis Pujol (Université Paris-Saclay). Available as a preprint: [68]

7.25. Axis 2: PAC-Bayesian Contrastive Unsupervised Representation Learning

Participant: Benjamin Guedj, Pascal Germain

Contrastive unsupervised representation learning (CURL) is the state-of-the-art technique to learn representations (as a set of features) from unlabelled data. While CURL has collected several empirical successes recently, theoretical understanding of its performance was still missing. In a recent work, Arora et al. (2019) provide the first generalisation bounds for CURL, relying on a Rademacher complexity. We extend their framework to the flexible PAC-Bayes setting, allowing to deal with the non-iid setting. We present PAC-Bayesian generalisation bounds for CURL, which are then used to derive a new representation learning algorithm. Numerical experiments on real-life datasets illustrate that our algorithm achieves competitive accuracy, and yields generalisation bounds with non-vacuous values.

Joint work with Kento Nozawa (University of Tokyo & RIKEN). Available as a preprint: [71]

7.26. Axis 2: Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly

Participant: Benjamin Guedj

When confronted with massive data streams, summarizing data with dimension reduction methods such as PCA raises theoretical and algorithmic pitfalls. Principal curves act as a nonlinear generalization of PCA and the present paper proposes a novel algorithm to automatically and sequentially learn principal curves from data streams. We show that our procedure is supported by regret bounds with optimal sublinear remainder terms. A greedy local search implementation (called `slpc`, for Sequential Learning Principal Curves) that incorporates both sleeping experts and multi-armed bandit ingredients is presented, along with its regret computation and performance on synthetic and real-life data.

Joint work with Le Li (Université d'Angers & iAdvize). Available as a preprint: [67]

7.27. Axis 2: PAC-Bayes Un-Expected Bernstein Inequality

Participant: Benjamin Guedj

We present a new PAC-Bayesian generalization bound. Standard bounds contain a $\sqrt{L_n \cdot KL/n}$ complexity term which dominates unless L_n, the empirical error of the learning algorithm's randomized predictions, vanishes. We manage to replace L_n by a term which vanishes in many more situations, essentially whenever the employed learning algorithm is sufficiently stable on the dataset at hand. Our new bound consistently beats state-of-the-art bounds both on a toy example and on UCI datasets (with large enough n). Theoretically, unlike existing bounds, our new bound can be expected to converge to 0 faster whenever a Bernstein/Tsybakov condition holds, thus connecting PAC-Bayesian generalization and *excess risk* bounds—for the latter it has long been known that faster convergence can be obtained under Bernstein conditions. Our main technical tool is a new concentration inequality which is like Bernstein's but with X^2 taken outside its expectation.

Joint work with Peter Grünwald (CWI), Zakaria Mhammedi (Australian National University).

This work has been accepted at NeurIPS 2019, will be presented as a poster in the main conference and as a oral in the workshop "Machine Learning with guarantees", and is included in the proceedings of NeurIPS 2019.

Published: [37]

7.28. Axis 2: Attributing and Referencing (Research) Software: Best Practices and Outlook from Inria

Participant: Benjamin Guedj

Software is a fundamental pillar of modern scientiic research, not only in computer science, but actually across all elds and disciplines. However, there is a lack of adequate means to cite and reference software, for many reasons. An obvious rst reason is software authorship, which can range from a single developer to a whole team, and can even vary in time. The panorama is even more complex than that, because many roles can be involved in software development: software architect, coder, debugger, tester, team manager, and so on. Arguably, the researchers who have invented the key algorithms underlying the software can also claim a part of the authorship. And there are many other reasons that make this issue complex. We provide in this paper a contribution to the ongoing eeorts to develop proper guidelines and recommendations for software citation, building upon the internal experience of Inria, the French research institute for digital sciences. As a central contribution, we make three key recommendations. (1) We propose a richer taxonomy for software contributions with a qualitative scale. (2) We claim that it is essential to put the human at the heart of the evaluation. And (3) we propose to distinguish citation from reference.

Joint work with Pierre Alliez, Roberto Di Cosmo, Alain Girault, Mohand-Said Hacid, Arnaud Legrand, Nicolas Rougier (Inria).

This work has been published in the journal Computing in Science and Engineering.

Published: [14]

7.29. Axis 2: Revisiting clustering as matrix factorisation on the Stiefel manifold

Participant: Benjamin Guedj

This paper studies clustering for possibly high dimensional data (*e.g.* images, time series, gene expression data, and many other settings), and rephrase it as low rank matrix estimation in the PAC-Bayesian framework. Our approach leverages the well known Burer-Monteiro factorisation strategy from large scale optimisation, in the context of low rank estimation. Moreover, our Burer-Monteiro factors are shown to lie on a Stiefel manifold. We propose a new generalized Bayesian estimator for this problem and prove novel prediction bounds for clustering. We also devise a componentwise Langevin sampler on the Stiefel manifold to compute this estimator.

Joint work with Stéphane Chrétien (Université Lyon-2). Available as a preprint: [55]

7.30. Axis 2: A Primer on PAC-Bayesian Learning

Participant: Benjamin Guedj

This survey on PAC-Bayesian learning has been the backbone to a successful proposal for an ICML 2019 plenary tutorial.

Generalised Bayesian learning algorithms are increasingly popular in machine learning, due to their PAC generalisation properties and flexibility. The present paper aims at providing a self-contained survey on the resulting PAC-Bayes framework and some of its main theoretical and algorithmic developments.

This work has been published in the proceedings of the French Mathematical Society. Published as [66].

7.31. Axis 2: Perturbed Model Validation: A New Framework to Validate Model Relevance

Participant: Benjamin Guedj

This paper introduces Perturbed Model Validation (PMV), a new technique to validate model relevance and detect overfitting or underfitting. PMV operates by injecting noise to the training data, re-training the model against the perturbed data, then using the training accuracy decrease rate to assess model relevance. A larger decrease rate indicates better concept-hypothesis fit. We realise PMV by perturbing labels to inject noise, and evaluate PMV on four real-world datasets (breast cancer, adult, connect-4, and MNIST) and nine synthetic datasets in the classification setting. The results reveal that PMV selects models more precisely and in a more stable way than cross-validation, and effectively detects both overfitting and underfitting.

It is a joint work with Jie Zhang, Earl Barr, John Shawe-Taylor (all with UCL), and Mark Harman (UCL & Facebook). Available as a preprint: [75].

7.32. Axis 2: Decentralized learning with budgeted network load using Gaussian copulas and classifier ensembles

Participant: Benjamin Guedj

We examine a network of learners which address the same classification task but must learn from different data sets. The learners cannot share data but instead share their models. Models are shared only one time so as to preserve the network load. We introduce DELCO (standing for Decentralized Ensemble Learning with COpulas), a new approach allowing to aggregate the predictions of the classifiers trained by each learner. The proposed method aggregates the base classifiers using a probabilistic model relying on Gaussian copulas. Experiments on logistic regressor ensembles demonstrate competing accuracy and increased robustness in case of dependent classifiers. A companion python implementation is available online.

Joint work with John Klein, Olivier Colot, Mahmoud Albardan (Université de Lille).

This work has been published in the proceedings of ECML-PKDD 2019, as part (oral presentation) of the workshop Decentralized Machine Learning at the Edge. Published: [34]

7.33. Axis 2: Online k-means Clustering

Participant: Benjamin Guedj

We study the problem of online clustering where a clustering algorithm has to assign a new point that arrives to one of k clusters. The specific formulation we use is the k-means objective: At each time step the algorithm has to maintain a set of k candidate centers and the loss incurred is the squared distance between the new point and the closest center. The goal is to minimize regret with respect to the best solution to the k-means objective ($\mathcal{C}$) in hindsight. We show that provided the data lies in a bounded region, an implementation of the Multiplicative Weights Update Algorithm (*MWUA*) using a discretized grid achieves a regret bound of $\widetilde{O}(\sqrt{T})$ in expectation. We also present an online-to-offline reduction that shows that an efficient no-regret online algorithm (despite being allowed to choose a different set of candidate centres at each round) implies an offline efficient algorithm for the k-means problem. In light of this hardness, we consider the slightly weaker requirement of comparing regret with respect to $(1+\epsilon)\mathcal{C}$ and present a no-regret algorithm with runtime $O\left(T(\mathrm{poly}(\log(T), k, d, 1/\epsilon)^{k(d+O(1))}\right)$. Our algorithm is based on maintaining an incremental coreset and an adaptive variant of the *MWUA*. We show that naïve online algorithms, such as *Follow The Leader*, fail to produce sublinear regret in the worst case. We also report preliminary experiments with synthetic and real-world data.

Joint work with Varun Kanade, Guy Rom (University of Oxford), Vincent Cohen-Addad (CNRS). Available as a preprint: [57]

7.34. Axis 2: Non-linear aggregation of filters to improve image denoising

Participant: Benjamin Guedj

We introduce a novel aggregation method to efficiently perform image denoising. Preliminary filters are aggregated in a non-linear fashion, using a new metric of pixel proximity based on how the pool of filters reaches a consensus. We provide a theoretical bound to support our aggregation scheme, its numerical performance is illustrated and we show that the aggregate significantly outperforms each of the preliminary filters.

Joint work with Juliette Rengot (Ecole des Ponts).

This work has been accepted at the Computing Conference 2020 (July 2020, London, UK) and will be included in the proceedings. Published: [33]

7.35. Axis 2: Multiple change-points detection with reproducing kernels

Participant: Alain Celisse

We tackle the change-point problem with data belonging to a general set. We build a penalty for choosing the number of change-points in the kernel-based method of Harchaoui and Cappé (2007). This penalty generalizes the one proposed by Lebarbier (2005) for a one-dimensional signal changing only through its mean. We prove a non-asymptotic oracle inequality for the proposed method, thanks to a new concentration result for some function of Hilbert-space valued random variables. Experiments on synthetic and real data illustrate the accuracy of our method, showing that it can detect changes in the whole distribution of data, even when the mean and variance are constant.

Joint work with Sylvain Arlot (Orsay) and Zaïd Harchaoui (Seattle). This work has been accepted in JMLR [15].

7.36. Axis 2: Analysis of early stopping rules based on discrepancy principle

Participant: Alain Celisse

We describe a general unified framework for analyzing the statistical performance of early stopping rules based on the minimum discrepancy principle (DP). Finite-sample bounds such as deviation or oracle inequalities are derived with high probability. Since it turns out that DP suffers some deficiencies when estimating smooth functions, refinements involving smoothing of the residuals are introduced and analyzed. Theoretical bounds established in the fixed design setting under mild assumptions such as the boundedness of the kernel. When focusing on the smoothed discrepancy principle, such bounds are even extended to the random design setting by means of a new change-of-norm argument

Joint work with Markus Reiß(Humboldt) and Martin Wahl (Humboldt). This work has been already presented several times in seminars.

7.37. Axis 3: Short-term air temperature forecasting using Nonparametric Functional Data Analysis and SARMA models

Participant: Sophie Dabo-Niang

Air temperature is a significant meteorological variable that affects social activities and economic sectors. In this paper, a non-parametric and a parametric approach are used to forecast hourly air temperature up to 24 h in advance. The former is a regression model in the Functional Data Analysis framework. The nonlinear regression operator is estimated using a kernel function. The smoothing parameter is obtained by a cross-validation procedure and used for the selection of the optimal number of closest curves. The other method applied is a Seasonal Autoregressive Moving Average (SARMA) model, the order of which is determined by the Bayesian Information Criterion. The obtained forecasts are combined using weights calculated based on the forecast errors. The results show that SARMA has a better performance for the first 6 forecasted hours, after which the Non-Parametric Functional Data Analysis (NPFDA) model provides superior results. Forecast pooling improves the accuracy of the forecasts.

It is a joint work with Stelian Curceac (Rothamsted Research, UK) Camille Ternynck (CERIM, Université de Lille) Taha B.M.J. Ouarda (INRS, Québec, Canada) Fateh Chebana (INRS, Québec, Canada). This work has been published in the journal Environmental Modelling and Software [18].

7.38. Axis 3: Mathematical Modeling and Study of Random or Deterministic Phenomena

Participant: Sophie Dabo-Niang

In order to identify mathematical modeling (including functional data analysis) and interdisciplinary research issues in evolutionary biology, epidemiology, epistemology, environmental and social sciences encountered by researchers in Mayotte, the first international conference on mathematical modeling (CIMOM'18) was held in Dembéni, Mayotte, from November 15 to 17, 2018, at the Centre Universitaire de Formation et de Recherche. The objective was to focus on mathematical research with interdisciplinarity. This contribution is a book discusses key aspects of recent developments in applied mathematical analysis and modeling. It was written after the international conference on mathematical modeling in Mayotte, where a call for chapters of the book was made. They were written in the form of journal articles, with new results extending the talks given during the conference and were reviewed by independent reviewers and book publishers It highlights a wide range of applications in the fields of biological and environmental sciences, epidemiology and social perspectives. Each chapter examines selected research problems and presents a balanced mix of theory and applications on some selected topics. Particular emphasis is placed on presenting the fundamental developments in mathematical analysis and modeling and highlighting the latest developments in different fields of probability and statistics. The chapters are presented independently and contain enough references to allow the reader to explore the various topics presented.

It is a joint work with Solym Manou-Abi and Jean-Jacques Salone (University of Mayotte, France). This book is to appear Wiley (ISTE) [21].

7.39. Axis 3: Categorical functional data analysis

Participants: Cristian Preda, Quentin Grimonprez, Vincent Vandewalle.

The research on functional data analysis is very actual. The R package "fda" is the most famous one implementing methodology for functional data. To the best of our knowledge, and quite surprisingly, there is no recent researches devoted to categorical functional data despite its ability to model real situations in different fields of applications: health and medicine (status of a patient over time), economy (status of the market), sociology (evolution of social status), and so on. We have developed the methodology to visualize, do dimension reduction and extract feature from categorical functional data. For this, the *cfda* R package has been developed.

7.40. Axis 4: Proteomic signature of early death in heart failure patients

Participants: Guillemette Marot, Vincent Vandewalle.

Heart failure (HF) remains a main cause of mortality worldwide. Risk stratification of patients with systolic chronic HF is critical to identify those who may benefit from advanced HF therapies. The aim of this study is to identify plasmatic proteins that could predict the early death (within 3 years) of HF patients with reduced ejection fraction hospitalized in CHRU de Lille. In this framework, we have performed LASSO logistic regression to perform variable selection in order to select candidates protein to predict early death in HF patients. An article has been accepted in Scientific Reports [19].

This is a joint work with Marie Cuvelliez, Florence Pinet and Christophe Bauters from INSERM.

7.41. Axis 4: Statistical analysis of high-throughput proteomic data

Participants: Guillemette Marot, Vincent Vandewalle, Wilfried Heyse.

From March until August 2019, Guillemette Marot and Vincent Vandewalle have supervised the internship of Wilfried Heyse (Master 2 Ingénieurie de Sytèmes Numériques). The purpose of this internship was to identify new circulating biomarkers of left ventricular remodelling (LVR) in patients suffering from myocardial infarction (MI). The aim is to precisely identify earlier after MI the patients at high risk of developing LVR that is quantified by imaging one year after MI. For that purpose, high throughput proteomic approach was used. This technology allows the measurement of 5000 proteins simultaneously. In parallel to these measures corresponding to the concentration of a protein in a plasma sample collected from one patient at a specific time, echocardiographic and clinical information will be collected on each of the 200 patients. Several approaches

have been used to predict the LVR based on proteins measurements. In particular penalized regression such as LASSO and variable clustering. Wilfried Heyse has now started a Phd Thesis granted by INSERM and supervised by Christophe Bauters, Guillemette Marot and Vincent Vandewalle. One of the main challenge is to take into account the variations of the biomarkers according to the time (several measurement times), in order the improve the understanding of biological mechanisms involved on LVR.

This is a joint work with Florence Pinet and Christophe Bauters from INSERM.

7.42. Axis 4: Linking different kinds of Omics data through a model-based clustering approach

Participants: Guillemette Marot, Vincent Vandewalle, Wilfried Heyse.

In this work, a mixture model allowing for genes clustering using both microarray (continuous) and RNAseq (count) expression data is proposed. More generally, it answers the clustering of variables issue, when variables are of different kinds (continuous and discrete here). Variables describing the same gene are constrained to belong to the same cluster. This constraint allows us to obtain a model that links the microarray and RNAseq measurements without needing parametric constraints on the form of this link. The proposed approach has been illustrated on simulated data, as well as on real data from TCGA (The Cancer Genome Atlas). It has been presented in an international conference [49].

This is a joint work with Camille Ternynck from EA2694.

7.43. Axis 4: Real-time Audio Sources Classification

Participants: Christophe Biernacki, Maxime Baelde.

This work addresses the recurring challenge of real-time monophonic and polyphonic audio source classification. The whole power spectrum is directly involved in the proposed process, avoiding complex and hazardous traditional feature extraction. It is also a natural candidate for polyphonic events thanks to its additive property in such cases. The classification task is performed through a nonparametric kernel-based generative modeling of the power spectrum. Advantage of this model is twofold: it is almost hypothesis free and it allows to straightforwardly obtain the maximum a posteriori classification rule of online signals. Moreover it makes use of the monophonic dataset to build the polyphonic one. Then, to reach the real-time target, the complexity of the method can be tuned by using a standard hierarchical clustering preprocessing of sound models, revealing a particularly efficient computation time and classification accuracy trade-off. The proposed method reveals encouraging results both in monophonic and polyphonic classification tasks on benchmark and owned datasets, even in real-time situations. This method also has several advantages compared to the state-of-the-art methods include a reduced training time, no hyperparameters tuning, the ability to control the computation - accuracy trade-off and no training on already mixed sounds for polyphonic classification. This work is now published in an international journal [16] and Maxime Baelde defended his PhD thesis on this topic this year [11].

It is a joint work with Raphaël Greff, from the A-Volute company.

7.44. Axis 4: Matching of descriptors evolving over time

Participants: Christophe Biernacki, Anne-Lise Bedenel.

In the web domain, and in particular for insurance comparison, data constantly evolve, implying that it is difficult to directly exploit them. For example, to do a classification, performing standard learning processes require data descriptors equal for both learning and test samples. Indeed, for answering web surfer expectation, online forms whence data come from are regularly modified. So, features and data descriptors are also regularly modified. In this work, it is introduced a process to estimate and understand connections between transformed data descriptors. This estimated matching between descriptors will be a preliminary step before applying later classical learning methods. Anne-Lise Bedenel defended her PhD thesis on this topic this year [12].

It is a joint work with Laetitia Jourdan, from Université de Lille.

7.45. Axis 4: Supervised multivariate discretization and levels merging for logistic regression

Participants: Christophe Biernacki, Vincent Vandewalle, Adrien Ehrhardt.

For regulatory and interpretability reasons, the logistic regression is still widely used by financial institutions to learn the refunding probability of a loan given the applicants characteristics from historical data. Although logistic regression handles naturally both quantitative and qualitative data, three ad hoc pre-processing steps are usually performed: firstly, continuous features are discretized by assigning factor levels to predetermined intervals; secondly, qualitative features, if they take numerous values, are grouped; thirdly, interactions (products between two different features) are sparsely introduced. By reinterpreting these discretized (resp. grouped) features as latent variables and by modeling the conditional distribution of each of these latent variables given each original feature with a polytomous logistic link (resp. contingency table), a novel model-based resolution of the discretization problem is introduced. Estimation is performed via a Stochastic Expectation-Maximization (SEM) algorithm and a Gibbs sampler to find the best discretization (resp. grouping) scheme w.r.t. any classical logistic regression loss (AIC, BIC, test set AUC,...). For detecting interacting features, the same scheme is used by replacing the Gibbs sampler by a Metropolis-Hastings algorithm. The good performances of this approach are illustrated on simulated and real data from Credit Agricole Consumer Finance. Adrien Ehrhardt defended his PhD thesis on this topic this year [13]. A preprint is being finalized to be submitted to an international journal or conference [62].

This is a joint work with Philippe Heinrich from Université de Lille.

7.46. Axis 4: MASSICCC Platform for SaaS Software Availability

Participant: Christophe Biernacki.

MASSICCC is a demonstration platform giving access through a SaaS (service as a software) concept to data analysis libraries developed at Inria. It allows to obtain results either directly through a website specific display (specific and interactive visual outputs) or through an R data object download. It started in October 2015 for two years and is common to the Modal team (Inria Lille) and the Select team (Inria Saclay). In 2016, two packages have been integrated: Mixmod and MixtComp (see the specific section about MixtComp). In 2017, the BlockCluster package has been integrated and also a particular attention to provide meaningful graphical outputs (for Mixmod,MixtComp and BlockCluster) directly in the web platform itself has led to some specific developments. In 2019, MASSICCC has been presented to a workshop [39]. In addition, the Mixtcomp software is now available on the CRAN depository. Currently, a preprint for an international journal dedicated to software is also in progress.

The MASSICCC platform is available here in the web: https://massiccc.lille.inria.fr

7.47. Axis 4: Domain adaptation from a pre-trained source model

Participants: Christophe Biernacki, Pascal Germain, Luxin Zhang.

Traditional statistical learning paradigm assumes the consistency between train and test data distributions. This rarely holds in many real-life applications. The domain adaptation paradigm proposes a variety of techniques to overcome this issue. Most of the works in this area seek either for a latent space where source and target data share the same distribution, or for a transformation of the source distribution to match the target one. Both strategies require learning a model on the transformed source data. An original scenario is studied where one is given a model that has been constructed using expertise on the source data that is not accessible anymore. To use directly this model on target data, we propose to learn a transformation from the target domain to the source domain. Up to our knowledge, this is a new perspective on domain adaptation. This learning problem is introduced and formalized. We study the assumptions and the sufficient conditions mandatory to guarantee a good accuracy when using the source model directly on transformed target data. By pursuing this idea, a new domain adaptation method based on optimal transport is proposed. We experiment our method on a fraud detection problem.

Luxin Zhang begun his PhD thesis on this topic and presented this early result in an international conference [51].

It is a joint work with Yacine Kessaci, both from Worldline.

7.48. Axis 4: Reject Inference Methods in Credit Scoring: a rational review

Participants: Christophe Biernacki, Vincent Vandewalle, Adrien Ehrhardt.

The granting process of all credit institutions is based on the probability that the applicant will refund his/her loan given his/her characteristics. This probability also called score is learnt based on a dataset in which rejected applicants are de facto excluded. This implies that the population on which the score is used will be different from the learning population. Thus, this biased learning can have consequences on the scorecard's relevance. Many methods dubbed reject inference have been developed in order to try to exploit the data available from the rejected applicants to build the score. However most of these methods are considered from an empirical point of view, and there is some lack of formalization of the assumptions that are really made, and of the theoretical properties that can be expected. We propose a formalisation of these usually hidden assumptions for some of the most common reject inference methods, and we discuss the improvement that can be expected. These conclusions are illustrated on simulated data and on real data from Credit Agricole Consumer Finance (CACF), a major European loan issuer. Adrien Ehrhardt defended his PhD thesis on this topic this year [13]. A preprint is being finalized to be submitted to an international journal or conference.

This is a joint work with Philippe Heinrich from Université de Lille.

7.49. Other: Projection Under Pairwise Control

Participant: Christophe Biernacki.

Visualization of high-dimensional and possibly complex (non-continuous for instance) data onto a low-dimensional space may be difficult. Several projection methods have been already proposed for displaying such high-dimensional structures on a lower-dimensional space, but the information lost is not always easy to use. Here, a new projection paradigm is presented to describe a non-linear projection method that takes into account the projection quality of each projected point in the reduced space, this quality being directly available in the same scale as this reduced space. More specifically, this novel method allows a straightforward visualization data in R2 with a simple reading of the approximation quality, and provides then a novel variant of dimensionality reduction. This work is now under minor revision in an international journal [53].

It is a joint work with Hiba Alawieh and Nicolas Wicker, both from Université de Lille.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. COLAS company

Participant: Christophe Biernacki.

COLAS is a world leader in the construction and maintenance of transport infrastructure. This bilateral contract aims at classifying mixed data obtained with sensors coming from a study of the aging of road surfacing. The challenge is to deal with many missing (sensors failures) and correlated data (sensors proximity).

8.2. Bilateral Grants with Industry

8.2.1. EIT-Sysbooster: Nokia - Apsys/Airbus

Participant: Alain Celisse.

Nokia and Airbus are two worldwide known companies respectively working in communications and transport areas. The purpose of this contract is to perform root cause analysis to reduce (at the end) the number of failures.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. ONCOLille partnership

Participants: Sophie Dabo-Niang, Cristian Preda.

ONCOLille is a regional scientific interest group whose purpose is to develop fundamental, translational (pre-clinical) and clinical interdisciplinary cancer research, particularly in the field of resistance to therapies. Sophie Dabo-Niang is member of the executive group.

9.2. National Initiatives

9.2.1. Programme of Investments for the Future (PIA)

Bilille is a member of the PIA "Infrastructures en biologie-santé" IFB, French Institute of Bioinformatics (https://www.france-bioinformatique.fr/en). As the co-head of the platform, Guillemette Marot is thus involved in this network.

9.2.2. RHU PreciNASH

Participant: Guillemette Marot.

RHU PreciNASH

Acronym: PreciNASH

Project title: Non-alcoholic steato-hepatitis (NASH) from disease stratification to novel therapeutic approaches

Coordinator: F. Pattou

Duration: 5 years

Partners: FHU Integra and Sanofi

Abstract: PreciNASH, project coordinated by Pr. F. Pattou (UMR 859, EGID), aims at better understanding non alcoholic stratohepatitis (NASH) and improving its diagnosis and care. In this RHU, Guillemette Marot supervises a 2 years post-doc, as her team EA 2694 is a member of the FHU Integra. EA 2694 is involved in the WP1 for the development of a clinical-biological model for the prediction of NASH. Other partners of the FHU are UMR 859, UMR 1011 and UMR 8199, these last three teams being part of the labex EGID (European Genomic Institute for Diabetes). Sanofi is the main industrial partner of the RHU PreciNASH. The whole project will last 5 years (2016-2021).

9.2.3. CNRS PEPS Blanc – BayesRealForRNN project

Participants: Pascal Germain, Vera Shalaeva.

BayesRealForRNN project: PAC-Bayesian theory for recurrent neural networks: a control theoretic approach

Coordinator: Mihaly Petreczky, CNRS, UMR 9189 CRIStAL, Université de Lille

Year: 2019

Abstract: The project proposes to analyze the mathematical correctness of deep learning algorithms by combining techniques from control theory and PAC-Bayesian statistical theory. More precisely, the project proposes to concentrate on recurrent neural networks (RNNs), develop their structure theory using techniques from contol theory, and then apply this structure theory to derive PAC-Bayesian error bounds for RNNs.

9.2.4. CNRS AMIES PEPS 2 - DiagChange project

Participants: Cristian Preda, Quentin Grimonprez.

DiagChange

Coordinator: Cristian Preda, Inria MODAL

Year: 2019

Abstract: The project proposes to study the topic of change detection distribution for multivariate signal in a industrial context.The project is in collaboration with the Diagrams start-up.

9.2.5. AMIES PEPS 1 - CADIS2

Participants: Serge Iovleff, Sophie Dabo-Niang, Cristian Preda.

Partners: Société SIRS https://www.sirs-fr.com/sirs/fr/

Acronym: CADIS2

Project title: Classification Automatique D'Images Sentinel-2

Coordinator: Serge Iovleff

Year: 2019

Duration: 1 year

Abstract: In the context of several European projects, SIRS is in charge of exploring the improvements to be made to the "High Resolution Layers" as well as future prototypes such as "CORINE Land Cover +", on a European scale using the Sentinel-2 images, through the project H2020 "ECo-LaSS". The CADIS2 project aims to develop, study and implement supervised classification methods to classify trees in predefined forest areas by SIRS.

9.2.6. AMIES PEPS 2 - MadiPa

Participants: Stéphane Girard, Serge Iovleff.

Partners: Société Phimeca http://phimeca.com/, Mistis team Inria Grenoble Rhône-Alpes

Acronym: MadiPa

Project title: Modèles Auto-associatifs pour la Dispersion de Polluants dans l'Atmosphère

Coordinator: Stéphane Iovleff

Duration: 18 month (start in december 2019)

Abstract: Our goal is to develop a method for predicting the dispersion of pollutants in the atmosphere from an initial emission map and meteorological data. A map of the probabilities of exceeding a critical threshold of pollutants will be estimated thanks to the construction of a meta-model: the large dimension of the problem is reduced by the use of auto-associative models, a non-linear extension of the Principal Components Analysis.

9.2.7. ANR

9.2.7.1. ANR APRIORI

Participants: Benjamin Guedj, Pascal Germain, Hemant Tyagi, Vera Shalaeva.

APRIORI 2019–2023, ANR PRC

PAC-Bayesian theory and algorithms for deep learning and representation learning.

Main coordinator of the project: Emilie Morvant, Université Jean Monnet.

Funding: 300k EUR.

2 partners - MODAL (Inria LNE), Hubert Curien Lab. (UMR CNRS 5516).

9.2.7.2. ANR BEAGLE

Participants: Benjamin Guedj, Pascal Germain.

BEAGLE 2019–2023, ANR JCJC

PAC-Bayesian theory and algorithms for agnostic learning

Main coordinator of the project: Benjamin Guedj

Funding: 180k EUR

The consortium also includes Pierre Alquier (RIKEN AIP, Japan), Peter Grünwald (CWI, The Netherlands), Rémi Bardenet (UMR CRIStAL 9189).

9.2.7.3. ANR SMILE

Participants: Christophe Biernacki, Vincent Vandewalle.

SMILE Project-2018-2022

ANR project (ANR SMILE - Statistical Modeling and Inference for unsupervised Learning at LargE-Scale)

Main coordinator of the project: Faicel Chamroukhi, LMNO, Université de Caen

4 partners - MODAL (Inria LNE), LMNO UMR CNRS 6139 (Caen), LMRS UMR CNRS 6085 (Rouen), LIS UMR CNRS 7020 (Toulon).

9.2.7.4. ANR TheraSCUD2022

Participant: Guillemette Marot.

Acronym: TheraSCUD2022

Project title: Targeting the IL-20/IL-22 balance to restore pulmonary, intestinal and metabolic homeostasis after cigarette smoking and unhealthy diet

Coordinator: P. Gosset

Duration: 3 years (2017-2020)

Partners: CIIL Institut Pasteur de Lille and UMR 1019 INRA Clermont-Ferrand

Abstract: TheraSCUD2022, project coordinated by P. Gosset (Institut Pasteur de Lille), studies inflammatory disorders associated with cigarette smoking and unhealthy diet (SCUD). Guillemette Marot is involved in this ANR project as head of bilille platform, and will supervise 1 year engineer on integration of omic data. The duration of this project is 3 years (2017-2020).

9.2.8. Working groups

Sophie Dabo-Niang belongs to the following working groups:

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

Benjamin Guedj belongs to the following working groups (GdR) of CNRS:

- ISIS (local referee for Inria Lille - Nord Europe)
- MaDICS
- MASCOT-NUM (local referee for Inria Lille - Nord Europe).

Guillemette Marot belongs to the StatOmique working group.

9.2.9. Other initiatives

Participants: Serge Iovleff, Cristian Preda, Vincent Vandewalle.

Serge Iovleff is the head of the project CloHe granted in 2016 by the Mastodons CNRS challenge "Big data and data quality". The project is axed on the design of classification and clustering algorithms for mixed data with missing values with applications to high spatial resolution multispectral satellite image time-series. Website. Cristian Preda and Vincent Vandewalle are also members of the CloHe project.

9.3. European Initiatives

9.3.1. FP7 & H2020 Projects

PERF-AI project (Nov 2018 - Nov 2020, involving Benjamin Guedj, Vincent Vandewalle - hired Florent Dewez, Arthur Taelpert). Two partners: Inria LNE and the company Safety Line (Paris, France).

Commercial aviation is already responsible for 3% of the total CO2 emissions, and with a constant growth rate of 5% per year, traffic will double within the next decade. With the support of new technologies such as Big Data, Artificial Intelligence, in-flight connectivity, major improvements can be introduced to optimize flight trajectories. PERF-AI focuses on the challenge of minimizing fuel consumption throughout the flight. The aim of PERF-AI is to provide a flight trajectory optimization prototype that implements new machine learning performance models.

The first step of the project that was carried out in the first year was to define, implement and test narrow system identification techniques. Several Machine Learning methods have been tried and have provided very encouraging initial results.

PERF-AI main objective is to provide a computation engine that can be used in two ways:

- support update of FMS that integrate individual aircraft performance models, that allow to perform accurate trajectory prediction;
- perform trajectory optimization on the ground using most accurate aircraft performance models.

9.3.2. Collaborations with Major European Organizations

Sophie Dabo-Niang is chair of EMS-CDC (European Mathematical society-Committe of Developing Countries).

Sophie Dabo-Niang is a member of the executive committee of CIMPA (International Centre of Pure and Applied Mathematics)

9.4. International Initiatives

9.4.1. Inria International Labs

9.4.1.1. 6PAC (IIL CWI-Inria)

Scientific leaders: Benjamin Guedj, Peter Grünwald.

Other members: Emilie Kaufmann (Inria LNE, EPI SequeL), Wouter Koolen (CWI).

Title: Making Probably Approximately Correct Learning Active, Sequential, Structure-aware, Efficient, Ideal and Safe

International Partner (Institution - Laboratory - Researcher):

CWI (Netherlands) - Machine Learning Group - Peter Grünwald (head)

Start year: 2018, renewed for 2019 and 2020

Webpage: https://bguedj.github.io/6pac/index.html

This project roots in statistical learning theory, which can be viewed as the theoretical foundations of machine learning. The most common framework is a setup in which one is given n training examples, and the goal is to build a predictor that would be efficient on new (similar) data. This efficiency should be supported by PAC (Probably Approximately Correct) guarantees, e.g. upper bounds on the excess risk of a predictor that hold with high probability. Such guarantees however often hold under stringent assumptions which are typically never met in real-life application, e.g., independent, identically distributed data. More realistic modelling of data has triggered many research efforts in several directions: first, accommodating possible data (e.g., dependent, heavy-tailed), and second, in the direction of sequential learning, in which the predictor can be built on the fly, while new data is gathered. We believe that an ever more realistic paradigm is active learning, a setup in which the learner actively requests data (possibly facing constraints, such as storage, velocity, cost, etc.)

and adapts its queries to optimize its performance. The 3-years objective of 6PAC (where 6 stands for Sequential, Active, Efficient, Structured, Ideal, Safe - the six research directions we intend to contribute to) is to pave the way to new PAC generalization and sample-complexity upper and lower bounds beyond batch learning. Our ambition is to contribute to several learning setups, ranging from sequential learning (where data streams are collected) to adaptive and active learning (where data streams are requested by the learning algorithm).

9.4.2. Inria International Partners

9.4.2.1. Declared Inria International Partners

A byproduct of Benjamin Guedj's sabbatical position at University College London (UCL) since Dec 2018 is a strengthened link between UCL and Inria. DGDS has established contact with UCL President in April 2019 and a MoU has been signed between UCL and Inria in December 2019. A research group (known as Inria@UCL) has been established by Benjamin Guedj within UCL, Department for Computer Science, Centre for Artificial Intelligence. Inria@UCL initiative is expected to grow in 2020 and possibly evolve into a joint team or more. A strategic partnership between Inria and UCL will be explored in 2020.

SIMERGE

Title: Statistical Inference for the Management of Extreme Risks and Global Epidemiology

International Partner (Institution - Laboratory - Researcher):

UGB (Senegal) - LERSTAD - Abdou Ka Diongue

Serge Iovleff and Sophie Dabo-Niang are associated members of SIMERGE.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

- Mihai Cucuringu (University of Oxford) visited Hemant Tyagi in January 2019 for a research visit of 1 week.
- Martin Wahl (Humboldt Universität from Berlin) visited Alain Celisse in March 2019 for a research visit of 1 week and November 2019 for a research visit of 1 week.
- Apoorv Vikram Singh is currently visiting Hemant Tyagi to work on a research project which is jointly supervised by Hemant Tyagi and Mihai Cucuringu (University of Oxford). The duration of the visit is 4 months (October 1, 2019 - January 31, 2020) and is partly funded by the Alan Turing Institute, London.
- Abdou Kâ Diongue visited Serge Iovleff in June 2019 for one month.

9.5.2. Visits to International Teams

9.5.2.1. Sabbatical programme

Since Dec 2018, Benjamin Guedj is on sabbatical at University College London (UCL). He is a PI of the UCL Centre for Artificial Intelligence (UCL AI) and a visiting researcher at the Alan Turing Institute. This has led to the Inria@UCL initiative, see supra.

9.5.2.2. Research Stays Abroad

- Sophie Dabo-Niang has visited University of Kuala Lumpur, Malaysia in August 2019 and University of Mohamed V, Morroco in December 2019.
- Serge Iovleff has visited University Gaston Berger, Senegal in February 2019 and gave a course entitled "Introduction to Statistical Learning".
- Hemant Tyagi visited Mihai Cucuringu and Benjamin Guedj at the Alan Turing Institute, UK from in October 2019.
- Alain Celisse visited Markus Reißand Martin Wahl at the Humboldt Universität, Germany in March and December 2019.
- Alain Celisse visited Benjamin Guedj at the University College London, UK in February-March and July-August 2019.
- Pascal Germain visited Benjamin Guedj at University College London, UK on several occasions totalling about 1.5 month in 2019.
- Cristian Preda visited Amarioarei Alexandru at University of Bucharest on several occasions totalling about 1 week in 2019.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Organisation

10.1.1.1. General Chair, Scientific Chair

- Hemant Tyagi and Pascal Germain are the organizers of the MODAL team scientific seminar.
- Sophie Dabo-Niang is the co-organizer of:
  - Session "Recent non-parametric approaches: Applications to envi- ronmental, hydrological, oceanological and economic data analyzes" of the ISI conference (62nd World Statistics Congress), 18-23, August, 2019, Kuala Lumpur, Malaysia.
  - Session "Modeling dependence through graphical models", CMStatistics, 14-16 December 2019, London.
- Christophe Biernacki is a co-organizer (and the chair) of the special session on co-clustering called "Co-clustering: model-based or model-free approaches" at the 62nd ISI World Statistics Congress 2019 in Malaysia [45]
- Alain Celisse is the organizer and the chair of two sessions on change-points detection and early stopping rules at ERCIM 2019, London.
- Vincent Vandewalle organized a session on Model-based and multivariate functional data at CRoNoS & MDA 2019.

10.1.2. Scientific Events: Selection

10.1.2.1. Member of the Conference Program Committees

- Sophie Dabo-Niang was a member of the scientific program committee of: CRONOS-MDA2019, 14-16 April 2019, Limassol, Cyprus and African Econometric Society Conference, 11-13 July 2019, Rabat, Morocco.
- Benjamin Guedj has been a PC member for UAI 2019 and IJCAI 2019.

10.1.2.2. Reviewer

- Hemant Tyagi acted as a reviewer for journals (Foundations of Computational Mathematics, Journal of Machine Learning Research, SIAM Journal on Scientific Computing, Advances in Data Analysis and Classification) and a conference (NeurIPS 2019).
- Pascal Germain acted as a reviewer for a journal (IEEE Transactions on Pattern Analysis and Machine Intelligence) and conferences (ICML 2019, ICLR 2019, NeurIPS 2019, COLT 2019, PHM 2019).
- Benjamin Guedj served as a reviewer for journals (JMLR, Neurocomputing, Journal of international research on robotics,...) and conferences (ICML 2019, ICLR 2019, NeurIPS 2019, COLT 2019, IJCAI 2019, UAI 2019, AISTATS 2019 & 2020).
- Alain Celisse acted as a reviewer for journals (IEEE Transactions on Pattern Analysis and Machine Intelligence, Annals of Statistics, JMLR,...) and conferences (ALT 2019, COLT 2019).
- Sophie Dabo-Niang acted as a reviewer for journals Annals of Statistics, JASA, Journal of Nonparametric statistics, TEST, METRIKA.
- Christophe Biernacki acted as a reviewer for CSDA, ESWA, JCGS, JMIV.
- Serge Iovleff acted as a reviewer for CSDA and Journal of Statistics and Computing.
- Vincent Vandewalle acted as a reviewer for ADAC, Statistics in Medicine, COST.
- Cristian Preda is a reviewer for Bernoulli, MCAP, ADAC.

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- Christophe Biernacki is an Associate Editor of the North-Western European Journal of Mathematics (NWEJM) and for Frontiers on the topic "Computational Methods for Data Analytics". He is also a Guest Editor for the Special Issue on Innovations in Model-Based Clustering and Classification of the journal Advances Data Analysis and Classification (ADAC).
- Cristian Preda is an associate editor for Methodological and Computing in applied probability (MCAP) and for Journal of Mathematical and Computer Science.
- Serge Iovleff is member of the Editorial Board of Astrostatistics (specialty section of Frontiers in Astronomy and Space Sciences)

10.1.4. Invited Talks

- Christophe Biernacki:
  - 3rd International Conference on Econometrics and Statistics (EcoSta 2019), Taiwan, June 25, 2019 [42]
  - 12th Scientific Meeting Classification and Data Analysis Group (CLADAG 2019), Italy, September 12, 2019 [41]
  - APSEM 2019 (Apprentissage et SEMantique) : écosystèmes pour la science ouverte et recherche par les données, Toulouse [39]
  - Séminary of Probability and Statistics of the University of Nice - Sophia Antopolis, France, November 19, 2019
- Hemant Tyagi:
  - Probability-Statistics seminar, Laboratoire de Mathematiques, Universite de Franche-Comte, Besancon, France, June 13, 2019
  - Probability-Statistics seminar, Université de Lille, France, September 18, 2019
- Pascal Germain:
  - Journées de la statistique, Nancy, France, June 6, 2019
- Benjamin Guedj:
  - Speaker (with John Shawe-Taylor) for a plenary tutorial at ICML 2019 (June 2019).

  - Several seminars in the UK (Gatsby unit, UCL CSML, UCL stats, ElementAI).
- Sophie Dabo-Niang:
  - 62nd World Statistics Congress, August, 18-23, 2019, Kuala Lumpur, Malaysia.
  - 7th Statistical Meeting of Avignon-Marseille, 14 June, 2019, Avignon, France.
  - Workshop on Non-Parametric Statistics, 18th, September, 2019, Grenoble, France.
  - CFIES 2019, 24-27 September 2019, Strasbourg, France.
  - Symposium of young Senegalese researchers in Pured and Applied Mathematics, December, 16-19, 2019, Mbour, Senegal.
- Cristian Preda:
  - Conference on scan statistics, IMS China, Dalian, 6-10 july, 2019.
  - Conference on categorical functional data at Romanian statistical society, 10-11 May 2019.
- Alain Celisse:
  - Session on Model selection at ERCIM 2018, London.
  - Several seminars in France (Nantes, Paris,...) and abroad WIAS Institute of Berlin.
- Vincent Vandewalle:
  - Session clustering categorical and mixed type data at IFCS 2019, Thessaloniki.

10.1.5. Leadership within the Scientific Community

- Benjamin Guedj and Pascal Germain are founding members of the Machine Learning and Artificial Intelligence group (MALIA) of the French statistical association (SFdS).
- Benjamin Guedj is the new French representative at the board of ECAS (since Nov 2019).
- Since 2017, Benjamin Guedj is an elected member of the board of the Statistical French Society (SFdS).

10.1.6. Scientific Expertise

Christophe Biernacki reviewed one project as an expert for the ANR and for Innoviris (Brussels).

10.1.7. Research Administration

- Sophie Dabo-Niang is the person in charge of the MeQAME axis of laboratory LEM, CNRS 9221.
- Christophe Biernacki is Scientific Head of the Inria Lille center since June 2017.
- Benjamin Guedj has been the list head for the election of the Evaluation Committee in June 2019. He has been an elected member of CE since 2017 and is a member of its executive board since Sept 2019.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

- Pascal Germain taught

  Master: Introduction aux réseaux de neurones, 15 heures, M2, Université de Lille, France
- Sophie Dabo-Niang is teaching

  Master: Spatial Statistics, 24h, M2, Université de Lille, France

  Master: Advanced Statistics, 24h, M2, Université de Lille, France

  Master: Multivariate Data Analyses, 24h, M2, Université de Lille, France

  Licence: Probability, 24h, L2, Université de Lille, France

  Licence: Multivariate Statistics, 24h, L3, Université de Lille, France
- Cristian Preda is teaching

Polytech'Lille engineer school: Linear Models, 48h.

Polytech'Lille engineer school: Advanced statistics, 48h.

Polytech'Lille engineer school: Biostatistics, 10h.

Polytech'Lille engineer school: Supervised clustering, 24h. France

- Christophe Biernacki is teaching

  New Master Data Science: Statistics, 24h, M1, Université de Lille, France
- Benjamin Guedj is teaching

  Advanced machine learning (M2, 6h), University College London, UK
- Serge Iovleff is teaching

  Licence: Analyse et méthodes numériques, 56h, Université de Lille, DUT Informatique

  Licence: R.O. et aide à la décision, 32h, Université de Lille, DUT Informatique
- Vincent Vandewalle is teaching

  Licence: Probability, 60h, Université de Lille, DUT STID

  Licence: Case study in statistics, 45h, Université de Lille, DUT STID

  Licence: R programming, 45h, Université de Lille, DUT STID

  Licence: Supervised clustering, 32h, Université de Lille, DUT STID

  Licence: Analysis, 24h, Université de Lille, DUT STID

10.2.2. Supervision

10.2.2.1. PhD defense:

- Maxime Baelde, September 20th 2019, supervised by Christophe Biernacki and Raphaël Greff on "Generative models for the classification and separation of real-time sound sources" [11].
- Anne-Lise Bedenel, April 1st 2019, supervised by Christophe Biernacki and Laetitia Jourdan on "Matching descriptors evolving over time: application to insurance comparison" [12].
- Adrien Ehrhardt, September 3rd 2019, supervised by Christophe Biernacki, Philippe Heinrich and Vincent Vandewalle on "Formalization and study of statistical problems in Credit Scoring" [13].

10.2.2.2. PhD in progress:

- Felix Biggs, Generative models and kernels, University College London, Sep 2019, Benjamin Guedj.
- Antoine Vendeville, Learning on graph to stop the propagation of fake news, University College London, Sep 2019, Benjamin Guedj.
- Luxin Zhang, Domain adaptation from a pre-trained source model – Application to fraud detection in electronic payments, February 2019, Christophe Biernacki, Pascal Germain, Yacine Kessac.
- Paul Viallard, Interpreting representation learning through PAC-Bayes theory, September 2019, Amaury Habrard, Emilie Morvant, Pascal Germain.
- Dang Khoi Pham, Planning and re-planning of nurses in an oncology department using a multi-objective and interdisciplinary approach, September 2016, Sophie Dabo-Niang.
- Solange Doumun, Performance evaluation and contribution to the development of multispectral image analysis strategies for automatic and rapid diagnosis of malaria, December 2018, Sophie Dabo-Niang.
- Alaa Ali Ayad, Statistical modeling of large spatial data and its applications in health, September 2018, Sophie Dabo-Niang.
- Wilfried Heyse, Prise en compte de la structure temporelle dans l'analyse statistique de données protéomiques à haut débit, October 2019, Christophe Bauters, Guillemette Marot and Vincent Vandewalle.

- Margot Selosse, October 2017, Christophe Biernacki and Julien Jacques.
- Filippo Antonazzo, October 2019, Christophe Biernacki and Christine Keribin.
- Yaroslav Averyanov, September 2016, Early stopping and filters estimators, Alain Celisse.

10.2.3. Juries

- Sophie Dabo-Niang acted as a reviewer and an examinator for PhD theses.
- Christophe Biernacki acted as a reviewer for the following PhD theses: Keefe Murphy December 17th 2019 (UCD Dublin), Jocelyn Chauvet April 19th 2019 (University of Montpellier).
- Christophe Biernacki acted as an examinator for the following HdR defenses: Madalina Olteanu December 10th 2019 (University Paris 1), Servane Gey February 7th 2019 (University Paris 5).
- Christophe Biernacki participated in the following juries of recruitment: MC jury université d'Avignon, Jury CR Lille, Jury DR Inria.
- Christophe Biernacki participated in the following juries of recruitment: MCF jury Université de Nice May 2019, Jury CR Paris 2019, Jury CR Inria (nationwide) 2019.
- Vincent Vandewalle participated in an MC jury Université Versailles Saint Quentin May 2019.
- Alain Celisse acted as a reviewer for the HdR defense of Zoltan Szabo in December 2019 at Ecole Polytechnique, Palaiseau.
- Cristian Preda acted as a referee for the PhD thesis of Amandine Schmutz, Université de Lyon 2, November 15, 2019.
- Cristian Preda acted as a referee for the HDR defense of Raluca Vernic, July 10, 2019, Universitatea Ovidiu, Constanta, Romania.

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

Christophe Biernacki acts as a member of the working group calculation for Inria.

10.3.2. Interventions

- Pascal Germain participated in the school activity "Jouer à débattre" of the organization "L'arbre des connaissances", as an expert in artificial intelligence (Lycée EIC de Tourcoing, February 7, 2019).
- Pascal Germain gave a vulgarization talk about neural networks and deep learning at the "Journée de l'Enseignement de l'Informatique et de l'Algorithmique" (Université de Lille, March 6, 2019).
- Christophe Biernacki participated in a round table "Deep Learning for PHM: Opportunities and Challenges", The 10th Prognostics and System Health Management Conference Paris, France, May 2 - May 5, 2019.
- Christophe Biernacki gave a talk at the Forum Teratec in June 12nd 2019, Ecole Polytechnique, Palaiseau with Margot Correard (DiagRAMS) on Predictive maintenance solution without additional sensors [40].
- Cristian Preda was invited as a speaker at IoT Week on predictive maintenance, Méaulte, December 4, 2019.
- Cristian Preda was invited as a speaker at IT Tour by Le Monde Informatique on artificial inteligence, November 14, 2019.

11. Bibliography

Major publications by the team in recent years

[1] P. ALQUIER, B. GUEDJ. *Simpler PAC-Bayesian Bounds for Hostile Data*, in "Machine Learning", 2018 [*DOI :* 10.1007/s10994-017-5690-0], https://hal.inria.fr/hal-01385064

[2] P. BATHIA, S. IOVLEFF, G. GOVAERT. *An R Package and C++ library for Latent block models: Theory, usage and applications*, in "Journal of Statistical Software", 2016, https://hal.archives-ouvertes.fr/hal-01285610

[3] C. BIERNACKI, A. LOURME. *Unifying Data Units and Models in (Co-)Clustering*, in "Advances in Data Analysis and Classification", May 2018, vol. 12, n° 41, https://hal.archives-ouvertes.fr/hal-01653881

[4] A. CELISSE. *Optimal cross-validation in density estimation with the L2-loss*, in "The Annals of Statistics", 2014, vol. 42, n° 5, pp. 1879–1910, https://hal.archives-ouvertes.fr/hal-00337058

[5] S. DABO-NIANG, C. TERNYNCK, A.-F. YAO. *Nonparametric prediction in the multivariate spatial context*, in "Journal of Nonparametric Statistics", 2016, vol. 28, n° 2, pp. 428-458 [*DOI :* 10.1080/10485252.2016.01.007], https://hal.inria.fr/hal-01425932

[6] J. DUBOIS, V. DUBOIS, H. DEHONDT, P. MAZROOEI, C. MAZUY, A. A. SÉRANDOUR, C. GHEERAERT, P. GUILLAUME, E. BAUGÉ, B. DERUDAS, N. HENNUYER, R. PAUMELLE, G. MAROT, J. S. CARROLL, M. LUPIEN, B. STAELS, P. LEFEBVRE, J. EECKHOUTE. *The logic of transcriptional regulator recruitment architecture at cis -regulatory modules controlling liver functions*, in "Genome Research", June 2017, vol. 27, n° 6, pp. 985–996 [*DOI :* 10.1101/GR.217075.116], https://hal.archives-ouvertes.fr/hal-01647846

[7] G. LETARTE, P. GERMAIN, B. GUEDJ, F. LAVIOLETTE. *Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks*, in "NeurIPS 2019", Vancouver, Canada, December 2019, https://hal.inria.fr/hal-02139432

[8] M. MARBAC, C. BIERNACKI, V. VANDEWALLE. *Model-based clustering of Gaussian copulas for mixed data*, in "Communications in Statistics - Theory and Methods", December 2016, https://hal.archives-ouvertes.fr/hal-00987760

[9] C. PREDA, A. DERMOUNE. *Parametrizations, fixed and random effects*, in "Journal of Multivariate Analysis", February 2017, vol. 154, pp. 162–176 [*DOI :* 10.1016/J.JMVA.2016.11.001], https://hal.archives-ouvertes.fr/hal-01655461

[10] H. TYAGI, J. VYBIRAL. *Learning general sparse additive models from point queries in high dimensions*, in "Constructive Approximation", January 2019, https://hal.inria.fr/hal-02379404

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] M. BAELDE. *Generative models for the classification and separation of real-time sound sources*, Université de Lille 1, September 2019, https://hal.archives-ouvertes.fr/tel-02399081

[12] A.-L. BEDENEL. *Matching descriptors evolving over time: application to insurance comparison*, Université de Lille I, April 2019, https://hal.archives-ouvertes.fr/tel-02399068

[13] A. EHRHARDT. *Formalization and study of statistical problems in Credit Scoring : Reject inference, discretization and pairwise interactions, logistic regression trees*, Université de Lille, September 2019, https://hal.archives-ouvertes.fr/tel-02302691

Articles in International Peer-Reviewed Journals

[14] P. ALLIEZ, R. DI COSMO, B. GUEDJ, A. GIRAULT, M.-S. HACID, A. LEGRAND, N. P. ROUGIER. *Attributing and Referencing (Research) Software: Best Practices and Outlook from Inria*, in "Computing in Science & Engineering", 2019, pp. 1-14, https://arxiv.org/abs/1905.11123 [*DOI :* 10.1109/MCSE.2019.2949413], https://hal.archives-ouvertes.fr/hal-02135891

[15] S. ARLOT, A. CELISSE, Z. HARCHAOUI. *A Kernel Multiple Change-point Algorithm via Model Selection*, in "Journal of Machine Learning Research", December 2019, vol. 20, n° 162, pp. 1–56, https://arxiv.org/abs/1202.3878 , https://hal.archives-ouvertes.fr/hal-00671174

[16] M. BAELDE, C. BIERNACKI, R. GREFF. *Real-Time Monophonic and Polyphonic Audio Classification from Power Spectra*, in "Pattern Recognition", August 2019, vol. 92, pp. 82-92 [*DOI :* 10.1016/J.PATCOG.2019.03.017], https://hal.archives-ouvertes.fr/hal-01834221

[17] M. BERNARDINI, A. BROSSA, G. CHINIGO, G. GROLEZ, G. TRIMAGLIO, L. ALLART, A. HULOT, G. MAROT, T. GENOVA, A. JOSHI, V. MATTOT, G. FROMONT, L. MUNARON, B. BUSSOLATI, N. PREVARSKAYA, A. FIORIO PLA, D. GKIKA. *Transient Receptor Potential Channel Expression Signatures in Tumor-Derived Endothelial Cells: Functional Roles in Prostate Cancer Angiogenesis*, in "Cancers", July 2019, vol. 11, n° 7, 956 p. [*DOI :* 10.3390/CANCERS11070956], https://hal.archives-ouvertes.fr/hal-02404061

[18] S. CURCEAC, C. TERNYNCK, T. B. OUARDA, F. CHEBANA, S. DABO-NIANG. *Short-term air temperature forecasting using Nonparametric Functional Data Analysis and SARMA models*, in "Environmental Modelling and Software", January 2019, vol. 111, pp. 394-408 [*DOI :* 10.1016/J.ENVSOFT.2018.09.017], https://hal.inria.fr/hal-01948928

[19] M. CUVELLIEZ, V. VANDEWALLE, M. BRUNIN, O. BESEME, A. HULOT, P. DE GROOTE, P. AMOUYEL, C. BAUTERS, G. MAROT, F. PINET. *Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction - Short title: Proteomic signature of early death in heart failure patients*, in "Scientific Reports", 2019, forthcoming, https://hal.inria.fr/hal-02400814

[20] M. CUVELLIEZ, V. VANDEWALLE, M. BRUNIN, O. BESEME, A. HULOT, P. DE GROOTE, P. AMOUYEL, C. BAUTERS, G. MAROT, F. PINET. *Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction*, in "Scientific Reports", December 2019, vol. 9, 19202 p. [*DOI :* 10.1038/S41598-019-55727-1], https://hal.archives-ouvertes.fr/hal-02414293

[21] S. DABO-NIANG, S. CURCEAC, C. TERNYNCK, T. B. OUARDA, F. CHEBANA, S. D. NIANG. *Short-term air temperature forecasting using Nonparametric Functional Data Analysis and SARMA models*, in "Environmental Modelling and Software", January 2019, vol. 111, pp. 394-408 [*DOI :* 10.1016/J.ENVSOFT.2018.09.017], https://hal.inria.fr/hal-02334991

[22] S. DABO-NIANG, B. THIAM. *Kernel regression estimation with errors-in-variables for random fields*, in "Afrika Matematika", 2019 [*DOI :* 10.1007/S13370-019-00654-7], https://hal.inria.fr/hal-02334993

[23] F. DEWEZ, V. MONTMIRAIL. *Decrypting the Hill Cipher via a Restricted Search over the Text-Space*, in "Linköping Electronic Conference Proceedings", June 2019, https://hal.univ-cotedazur.fr/hal-02271395

[24] A. EFTEKHARI, J. TANNER, A. THOMPSON, B. TOADER, H. TYAGI. *Sparse non-negative super-resolution — simplified and stabilised*, in "Applied and Computational Harmonic Analysis", August 2019 [*DOI :* 10.1016/J.ACHA.2019.08.004], https://hal.inria.fr/hal-02379445

[25] P. GERMAIN, A. HABRARD, F. LAVIOLETTE, E. MORVANT. *PAC-Bayes and Domain Adaptation*, in "Neurocomputing", 2020, vol. 379, pp. 379-397, https://arxiv.org/abs/1707.05712 [*DOI :* 10.1016/J.NEUCOM.2019.10.105], https://hal.archives-ouvertes.fr/hal-01563152

[26] A. GOYAL, E. MORVANT, P. GERMAIN, M.-R. AMINI. *Multiview Boosting by Controlling the Diversity and the Accuracy of View-specific Voters*, in "Neurocomputing", 2019, https://arxiv.org/abs/1808.05784 , forthcoming [*DOI :* 10.1016/J.NEUCOM.2019.04.072], https://hal.archives-ouvertes.fr/hal-01857463

[27] D. A. MOGILENKO, J. HAAS, L. L'HOMME, S. FLEURY, S. QUEMENER, M. LEVAVASSEUR, C. BECQUART, J. WARTELLE, A. BOGOMOLOVA, L. PINEAU, O. MOLENDI-COSTE, S. LANCEL, H. DEHONDT, C. GHEERAERT, A. MELCHIOR, C. DEWAS, A. NIKITIN, S. PIC, N. RABHI, J.-S. ANNICOTTE, S. OYADOMARI, T. VELASCO-HERNANDEZ, J. CAMMENGA, M. FORETZ, B. VIOLLET, M. VUKOVIC, A. VILLACRECES, K. KRANC, P. CARMELIET, G. MAROT, A. BOULTER, S. J. TAVERNIER, L. BEROD, M. P. LONGHI, C. PAGET, S. JANSSENS, D. STAUMONT-SALLÉ, E. AKSOY, B. STAELS, D. DOMBROWICZ. *Metabolic and innate immune cues merge into a specific inflammatory response via unfolded protein-response (UPR)*, in "Cell", May 2019, vol. 177, n^{o} 5, pp. 1201-1216.e19, Erratum in : Metabolic and Innate Immune Cues Merge into a Specific Inflammatory Response via the UPR. [Cell. 2019], forthcoming [*DOI :* 10.1016/J.CELL.2019.03.018], https://www.hal.inserm.fr/inserm-02084447

[28] M. SELOSSE, J. JACQUES, C. BIERNACKI, F. COUSSON-GÉLIE. *Analysing a quality of life survey using a co-clustering model for ordinal data and some dynamic implications*, in "Journal of the Royal Statistical Society: Series C Applied Statistics", July 2019, https://hal.archives-ouvertes.fr/hal-01643910

[29] M. SELOSSE, J. JACQUES, C. BIERNACKI. *Model-based co-clustering for mixed type data*, in "Computational Statistics and Data Analysis", 2020, vol. 144, 106866 p. [*DOI :* 10.1016/J.CSDA.2019.106866], https://hal.archives-ouvertes.fr/hal-01893457

[30] H. TYAGI, J. VYBIRAL. *Learning general sparse additive models from point queries in high dimensions*, in "Constructive Approximation", January 2019, https://hal.inria.fr/hal-02379404

Invited Conferences

[31] C. BIERNACKI, G. CELEUX, J. JOSSE, F. LAPORTE. *Dealing with missing data in model-based clustering through a MNAR model*, in "CRoNos & MDA 2019 - Meeting and Workshop on Multivariate Data Analysis and Software", Limassol, Cyprus, April 2019, https://hal.inria.fr/hal-02103347

International Conferences with Proceedings

[32] M. CUCURINGU, P. DAVIES, A. GLIELMO, H. TYAGI. *SPONGE: A generalized eigenproblem for clustering signed networks*, in "AISTATS", Okinawa, Japan, April 2019, https://hal.inria.fr/hal-02379505

[33] B. GUEDJ, J. RENGOT. *Non-linear aggregation of filters to improve image denoising*, in "Computing Conference 2020", London, United Kingdom, July 2020, https://hal.inria.fr/hal-02086856

[34] J. KLEIN, M. ALBARDAN, B. GUEDJ, O. COLOT. *Decentralized learning with budgeted network load using Gaussian copulas and classifier ensembles*, in "ECML-PKDD, Decentralized Machine Learning at

the Edge Workshop", Wurzburg, Germany, September 2019, https://arxiv.org/abs/1804.10028 , https://hal.archives-ouvertes.fr/hal-01779989

[35] G. LETARTE, P. GERMAIN, B. GUEDJ, F. LAVIOLETTE. *Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks*, in "NeurIPS 2019", Vancouver, Canada, December 2019, https://hal.inria.fr/hal-02139432

[36] G. LETARTE, E. MORVANT, P. GERMAIN. *Pseudo-Bayesian Learning with Kernel Fourier Transform as Prior*, in "The 22nd International Conference on Artificial Intelligence and Statistics", Naha, Japan, Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019,, 2019, https://arxiv.org/abs/1810.12683 , https://hal.archives-ouvertes.fr/hal-01908555

[37] Z. MHAMMEDI, P. GRÜNWALD, B. GUEDJ. *PAC-Bayes Un-Expected Bernstein Inequality*, in "NeurIPS 2019", Vancouver, Canada, December 2019, https://arxiv.org/abs/1905.13367 , https://hal.inria.fr/hal-02401295

[38] V. SHALAEVA, A. FAKHRIZADEH ESFAHANI, P. GERMAIN, M. PETRECZKY. *Improved PAC-Bayesian Bounds for Linear Regression*, in "Thirty-Fourth AAAI Conference on Artificial Intelligence", New York, United States, February 2020, https://arxiv.org/abs/1912.03036 , https://hal.inria.fr/hal-02396556

Conferences without Proceedings

[39] C. BIERNACKI. *MASSICCC: A SaaS Platform for Clustering and Co-Clustering of Mixed Data*, in "APSEM 2019 ((Apprentissage et SEMantique)) : éco-systèmes pour la science ouverte et recherche par les données", Toulouse, France, October 2019, https://hal.archives-ouvertes.fr/hal-02399180

[40] C. BIERNACKI, M. CORRÉARD. *Predictive maintenance solution without additional sensors*, in "Forum TERATEC", Palaiseau, France, June 2019, https://hal.archives-ouvertes.fr/hal-02399046

[41] C. BIERNACKI, A. LOURME. *Unifying Data Units and Models in (Co-)Clustering*, in "CLADAG 2019 - 12th Scientific Meeting Classification and Data Analysis Group", Cassino, Italy, September 2019, https://hal.archives-ouvertes.fr/hal-02398982

[42] C. BIERNACKI, M. MARBAC, V. VANDEWALLE. *Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering*, in "3rd International Conference on Econometrics and Statistics (EcoSta 2019)", Taichung, Taiwan, June 2019, https://hal.archives-ouvertes.fr/hal-02398999

[43] A. CONSTANTIN, M. FAUVEL, S. GIRARD, S. IOVLEFF, Y. TANGUY. *Classification de Signaux Multi-dimensionnels Irrégulièrement Echantillonnés*, in "2019 - Journée Jeunes Chercheurs MACLEAN du GDR MADICS", Paris, France, December 2019, pp. 1-2, https://hal.archives-ouvertes.fr/hal-02394120

[44] L. GAUTHERON, P. GERMAIN, A. HABRARD, G. LETARTE, E. MORVANT, M. SEBBAN, V. ZANTEDESCHI. *Revisite des "random Fourier features" basée sur l'apprentissage PAC-Bayésien via des points d'intérêts*, in "CAp 2019 - Conférence sur l'Apprentissage automatique", Toulouse, France, July 2019, https://hal.archives-ouvertes.fr/hal-02148600

[45] C. KERIBIN, C. BIERNACKI. *Co-clustering: model based or model free approaches*, in "ISI WSC 2019 - 62nd ISI World Statistics Congress", Kuala Lumpur, Malaysia, August 2019, https://hal.archives-ouvertes.fr/hal-02399031

[46] C. KERIBIN, C. BIERNACKI. *Le modèle des blocs latents, une méthode régularisée pour la classification en grande dimension*, in "JdS 2019 - 51èmes Journées de Statistique de la SFdS", Nancy, France, June 2019, https://hal.archives-ouvertes.fr/hal-02391379

[47] F. LAPORTE, C. BIERNACKI, G. CELEUX, J. JOSSE. *Modèles de classification non supervisée avec données manquantes non au hasard*, in "JdS 2019 - 51e journées de statistique de la Sfds", Nancy, France, June 2019, https://hal.archives-ouvertes.fr/hal-02398984

[48] M. MARBAC-LOURDELLE, C. BIERNACKI, V. VANDEWALLE. *Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering*, in "SPSR 2019", Bucarest, Romania, April 2020, https://hal.archives-ouvertes.fr/hal-02400486

[49] V. VANDEWALLE, C. TERNYNCK, G. MAROT. *Linking different kinds of Omics data through a model-based clustering approach*, in "IFCS 2019", Thessalonique, Greece, August 2019, https://hal.archives-ouvertes.fr/hal-02400525

[50] P. VIALLARD, R. EMONET, P. GERMAIN, A. HABRARD, E. MORVANT. *Interpreting Neural Networks as Majority Votes through the PAC-Bayesian Theory*, in "Workshop on Machine Learning with guarantees @ NeurIPS 2019", Vancouver, Canada, 2019, https://hal.archives-ouvertes.fr/hal-02335762

[51] L. ZHANG, C. BIERNACKI, P. GERMAIN, Y. KESSACI. *Domain Adaptation from a Pre-trained Source Model: Application on fraud detection tasks*, in "12th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2019)", London, United Kingdom, December 2019, https://hal.archives-ouvertes.fr/hal-02399003

Scientific Books (or Scientific Book chapters)

[52] S. DABO-NIANG, S. MANOU-ABI, S. JEAN-JACQUES. *Mathematical Modeling and Study of Random or Deterministic Phenomena,*, Wiley, 2020, https://hal.inria.fr/hal-02334997

Other Publications

[53] H. ALAWIEH, N. WICKER, C. BIERNACKI. *Projection under pairwise distance controls*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01420662

[54] C. BIERNACKI, M. MARBAC, V. VANDEWALLE. *Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01949155

[55] S. CHRÉTIEN, B. GUEDJ. *Revisiting clustering as matrix factorisation on the Stiefel manifold*, March 2019, https://arxiv.org/abs/1903.04479 - working paper or preprint, https://hal.inria.fr/hal-02064396

[56] S. CHRÉTIEN, H. TYAGI. *Multi-kernel unmixing and super-resolution using the Modified Matrix Pencil method*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02379598

[57] V. COHEN-ADDAD, B. GUEDJ, V. KANADE, G. ROM. *Online k-means Clustering*, December 2019, https://arxiv.org/abs/1909.06861 - 11 pages, 1 figure, https://hal.inria.fr/hal-02401290

[58] A. CONSTANTIN, M. FAUVEL, S. GIRARD, S. IOVLEFF. *Classification de Signaux Multidimensionnels Irrégulièrement Échantillonnés*, August 2019, GRETSI 2019 - 27e Colloque francophone de traitement du signal et des images, Poster, https://hal.archives-ouvertes.fr/hal-02276255

[59] A. CONSTANTIN, M. FAUVEL, S. GIRARD, S. IOVLEFF. *Supervised classification of multidimensional and irregularly sampled signals*, April 2019, 1 p. , Statlearn 2019 - Workshop on Challenging problems in Statistical Learning, Poster, https://hal.archives-ouvertes.fr/hal-02092347

[60] M. CUCURINGU, H. TYAGI. *Provably robust estimation of modulo 1 samples of a smooth function with applications to phase unwrapping*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02379573

[61] A. D'ASPREMONT, M. CUCURINGU, H. TYAGI. *Ranking and synchronization from pairwise measurements via SVD*, October 2019, https://arxiv.org/abs/1906.02746 - 42 pages, 9 figures, https://hal.archives-ouvertes.fr/hal-02340372

[62] A. EHRHARDT, C. BIERNACKI, V. VANDEWALLE, P. HEINRICH. *Feature quantization for parsimonious and interpretable predictive models*, March 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01949135

[63] M. P. B. GALLAUGHER, C. BIERNACKI, P. D. MCNICHOLAS. *Parameter-Wise Co-Clustering for High-Dimensional Data*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01862824

[64] L. GAUTHERON, P. GERMAIN, A. HABRARD, E. MORVANT, M. SEBBAN, V. ZANTEDESCHI. *Learning Landmark-Based Ensembles with Random Fourier Features and Gradient Boosting*, June 2019, https://arxiv.org/abs/1906.06203 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02148618

[65] B. GUEDJ, B. S. DESIKAN. *Kernel-Based Ensemble Learning in Python*, December 2019, https://arxiv.org/abs/1912.08311 - 11 pages, https://hal.inria.fr/hal-02443097

[66] B. GUEDJ. *A Primer on PAC-Bayesian Learning*, May 2019, working paper or preprint, https://hal.inria.fr/hal-01983732

[67] B. GUEDJ, L. LI. *Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly*, May 2019, working paper or preprint, https://hal.inria.fr/hal-01796011

[68] B. GUEDJ, L. PUJOL. *Still no free lunches: the price to pay for tighter PAC-Bayes bounds*, December 2019, https://arxiv.org/abs/1910.04460 - working paper or preprint, https://hal.inria.fr/hal-02401286

[69] F. LAPORTE, C. BIERNACKI, G. CELEUX, J. JOSSE. *Model-based clustering with missing not at random data. Missing mechanism*, July 2019, Working Group on Model-Based Clustering Summer Session, Poster, https://hal.archives-ouvertes.fr/hal-02398987

[70] G. MAZO, Y. AVERYANOV. *Constraining kernel estimators in semiparametric copula mixture models*, March 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01774629

[71] K. NOZAWA, P. GERMAIN, B. GUEDJ. *PAC-Bayesian Contrastive Unsupervised Representation Learning*, December 2019, https://arxiv.org/abs/1910.04464 - working paper or preprint, https://hal.inria.fr/hal-02401282

[72] M. SELOSSE, J. JACQUES, C. BIERNACKI. *Textual data summarization using the Self-Organized Co-Clustering model*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02115294

[73] M. SELOSSE, J. JACQUES, C. BIERNACKI. *ordinalClust: an R package for analyzing ordinal data*, December 2019, working paper or preprint, https://hal.inria.fr/hal-01678800

[74] S. N. SYLLA, S. DABO-NIANG, C. LOUCOUBAR. *Functional data analysis of parasite densities in the Senegalese villages of Dielmo and NDiop*, October 2019, working paper or preprint, https://hal.inria.fr/hal-02335001

[75] J. ZHANG, E. T. BARR, B. GUEDJ, M. HARMAN, J. SHAWE-TAYLOR. *Perturbed Model Validation: A New Framework to Validate Model Relevance*, May 2019, working paper or preprint, https://hal.inria.fr/hal-02139208