RESEARCH CENTRE

**Inria Center
at the University of Lille**

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2022
ACTIVITY REPORT

Project-Team
MODAL

MOdel for Data Analysis and Learning

IN COLLABORATION WITH: Laboratoire Paul Painlevé (LPP)

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Contents

Project-Team MODAL 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3
2.1 Context . . . 3
2.2 Goals . . . 4

3 Research program 4
3.1 Research axis 1: Unsupervised learning . . . 4
3.2 Research axis 2: Performance assessment . . . 4
3.3 Research axis 3: Functional data . . . 4
3.4 Research axis 4: Applications motivating research . . . 5

4 Application domains 5
4.1 Economic world . . . 5
4.2 Biology and health . . . 5

5 Social and environmental responsibility 5

6 Highlights of the year 5
6.1 Awards . . . 5

7 New software and platforms 6
7.1 New software . . . 6
7.1.1 MixtComp.V4 . . . 6
7.1.2 cfda . . . 6
7.1.3 ClusPred . . . 6
7.1.4 visCorVar . . . 7
7.1.5 metaRNASeq . . . 7
7.1.6 HDSpatialScan . . . 8
7.1.7 MLGL . . . 8
7.2 New platforms . . . 8
7.2.1 MASSICCC Platform . . . 8

8 New results 9
8.1 Axis 1: Co-clustering as a (very) parsimonious clustering . . . 9
8.2 Axis 1: Relaxing the identically distributed assumption in Gaussian co-clustering for high dimensional data . . . 9
8.3 Axis 1: Dealing with Missing Data in Model-based Clustering through a MNAR Model . . . 9
8.4 Axis 1: Predictive Clustering . . . 9
8.5 Axis 1: A Binned Technique for Scalable Model-based Clustering on Huge Datasets . . . 10
8.6 Axis 1: Forecasting elections results via the voter model with stubborn nodes . . . 10
8.7 Axis 1: Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly . . . 10
8.8 Axis 1&2: An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees . . . 11
8.9 Axis 1&2: Seeded graph matching for the correlated Wigner model via the projected power method . . . 11
8.10 Axis 1&2: Dynamic Ranking and Translation Synchronization . . . 11
8.11 Axis 1&2: Minimax Optimal Clustering of Bipartite Graphs with a Generalized Power Method 12
8.12 Axis 2: Asymptotic efficiency of some nonparametric tests for location on hyperspheres . . . 12
8.13 Axis 2: k-nearest neighbors prediction and classification for spatial data . . . 12
8.14 Axis 2: Progress in Self-Certified Neural Networks . . . 13
8.15 Axis 2: MMD Aggregated Two-Sample Test . . . 13
8.16 Axis 2: Learning PAC-Bayes Priors for Probabilistic Neural Networks . . . 13

8.17 Axis 2: On Margins and Derandomisation in PAC-Bayes . . . 14
8.18 Axis 2: Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound 14
8.19 Axis 2: Differentiable PAC–Bayes Objectives with Partially Aggregated Neural Networks . . 14
8.20 Axis 2: PAC-Bayes Unleashed: Generalisation Bounds with Unbounded Losses . . . 15
8.21 Axis 2: Still No Free Lunches: The Price to Pay for Tighter PAC-Bayes Bounds . . . 15
8.22 Axis 3: Non-parametric statistical analysis of spatially distributed functional data . . . 15
8.23 Axis 3: Clustering spatial functional data . . . 15
8.24 Axis 3: Regression models for spatially distributed autoregressive functional data . . . 16
8.25 Axis 3: Investigating spatial scan statistics for multivariate functional data . . . 16
8.26 Axis 3: PLS regression approach for multivariate functional data with different domains . . 16
8.27 Axis 3: Group lasso regression for spatially dependent functional data . . . 16
8.28 Axis 4: Statistical analysis of high-throughput proteomic data . . . 17
8.29 Axis 4: Multi-layer group Lasso . . . 17
8.30 Axis 4: Statistical analysis of transcriptomics data . . . 17
8.31 Axis 4: Statistical analysis of proteomic data with empirical bayesian approaches . . . 17
8.32 Axis 4: Multi-omics data analysis . . . 18
8.33 Axis 4: Interpretable Domain Adaptation for Hidden Subdomain Alignment in the Context of Pre-trained Source Models . . . 18
8.34 Axis 4: Interpretable Domain Adaptation Using Unsupervised Feature Selection on Pre-trained Source Models . . . 18
8.35 Axis 4: Single cell classification using statistical learning on mechanical properties measured by mems tweezers . . . 19
8.36 Axis 4: Dimensionality Reduction and Bandwidth Selection for Spatial Kernel Discriminant Analysis . . . 19
8.37 Axis 4: A kernel discriminant analysis for spatially dependent data . . . 20

9 Bilateral contracts and grants with industry **20**
9.1 Bilateral contracts with industry . . . 20
9.2 Bilateral grants with industry . . . 21

10 Partnerships and cooperations **22**
10.1 International initiatives . . . 22
10.1.1 Visits to international teams . . . 22
10.2 European initiatives . . . 23
10.2.1 H2020 projects . . . 23
10.3 National initiatives . . . 23
10.3.1 ANR . . . 25
10.3.2 RHU and FHU . . . 26
10.3.3 Working groups . . . 28
10.4 Regional initiatives . . . 28

11 Dissemination **29**
11.1 Promoting scientific activities . . . 29
11.1.1 Scientific events: organisation . . . 29
11.1.2 Scientific events: selection . . . 29
11.1.3 Journal . . . 29
11.1.4 Invited talks . . . 30
11.1.5 Leadership within the scientific community . . . 30
11.1.6 Scientific expertise . . . 30
11.1.7 Research administration . . . 31
11.2 Teaching - Supervision - Juries . . . 31
11.2.1 Teaching . . . 31
11.2.2 Supervision . . . 32
11.2.3 Juries . . . 32
11.3 Popularization . . . 33

11.3.1 Interventions . . . 33

12 Scientific production 33

12.1 Major publications . . . 33

12.2 Publications of the year . . . 33

12.3 Other . . . 36

12.4 Cited publications . . . 36

Project-Team MODAL

Creation of the Project-Team: 2012 January 01

Keywords

Computer sciences and digital sciences

A3.1.4. – Uncertain data

A3.1.10. – Heterogeneous 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

A5.9.2. – Estimation, modeling

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 members, visitors, external collaborators

Research Scientists

- Christophe Biernacki [INRIA, Senior Researcher, HDR]
- Benjamin Guedj [INRIA, Researcher]
- Hemant Tyagi [INRIA, Researcher]

Faculty Members

- Cristian Preda [Team leader, UNIV LILLE, Professor, HDR]
- Sophie Dabo [UNIV LILLE, Professor, HDR]
- Guillemette Marot [UNIV LILLE, Associate Professor, HDR]
- Vincent Vandewalle [Inria Lille, until Aug 2022, From September 1st, he is Professor at Université Cote d'Azur, HDR]

Post-Doctoral Fellows

- Ernesto Javier Araya Valdivia [INRIA]
- Rim Essifi [Inria, from Mar 2022]
- Valentina Zantedeschi [Inria, until Jul 2022]

PhD Students

- Reuben Adams [UCL]
- François Bassac [Décathlon, CIFRE, from Oct 2022]
- Felix Biggs [UCL]
- Clarisse Boinay [Seckiot, CIFRE]
- Guillaume Braun [INSEE]
- Théophile Cantelobre [Inria, until May 2022]
- Camille Frevent [UNIV LILLE]
- Maxime Haddouche [UNIV LILLE]
- Wilfried Heyse [INSERM, until Sep 2022]
- Eglantine Karle [INRIA]
- Etienne Kronert [WORDLINE]
- Issam Ali Moindjie [INRIA]
- Antoine Picard [UCL]
- Axel Potier [ADEO]
- Antonin Schrab [UCL]
- Antoine Vandeville [UCL]

Technical Staff

- Rachid Boulkhir [INRIA, Engineer, from Oct 2022]
- Ismat Chaib Draa [ALICANTE, Engineer]

Interns and Apprentices

- Myriam Benbahlouli [Saint-Gobain, Apprentice, until Sep 2022]
- Maxence Buisson [Polytech'Lille, from Jun 2022]
- Emile Moubarak [Ecole Centrale Lille, until Jun 2022]
- Whillem Tongo [Polytech'Lille, Intern, from May 2022 until Aug 2022]

Administrative Assistant

- Anne Rejl [INRIA]

Visiting Scientists

- Ayush Bhandari [IMPERIAL COLLEGE LDN, from Oct 2022 until Oct 2022]
- Stephane Chretien [UDL, from Jun 2022 until Jun 2022]
- Abderrazek Karoui [Université de Carthage, Tunisie, from Oct 2022 until Oct 2022]
- Mustapha Lebbah [UVSQ, from Sep 2022 until Sep 2022]
- Cathy Maugis-Rabusseau [INSA Toulouse, from Sep 2022 until Sep 2022]
- Allou Same [UNIV GUSTAVE EIFFEL, from Sep 2022 until Sep 2022]
- Cinzia Viroli [UNIV BOLOGNE, from Sep 2022 until Sep 2022]

External Collaborator

- Alain Celisse [UNIV PARIS I, HDR]

2 Overall objectives

2.1 Context

In several respects, modern society has strengthened the need for statistical analysis both from the applied and theoretical points of view. 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 of 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 to respectively as unsupervised learning or supervised learning, even if it is not limited to them or other names exist depending on communities. Somehow, 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 "manual" 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 etc.) 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 some projects treated by 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 etc. 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 etc.). Time series are an emblematic example even if it should not be limited to them (spectral data, spatial data etc.). 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 etc.). 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 its 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, Safety Line (through the PERF-AI consortium), Agence d'Urbanisme Métropole Européenne de Lille, ASYGN SAS (MEMs, joint Cytomems ANR project), HORIBA France SAS (Raman spectrometry)

4.2 Biology and health

The second main application domain of the team is biology and health. Members of the team are involved in the supervision and scientific animation of bilille, the bioinformatics platform of Lille, and of OncoLille Institute. Members of the team also co-supervise PhD students of Inserm teams.

5 Social and environmental responsibility

MODAL has not any social and environmental responsibility.

6 Highlights of the year

6.1 Awards

- Christophe Biernacki was elected as a Vice-head of the SFdS (Société Française de Statistique) since June 2022, which is the French society specialized in statistics, whose mission is to promote the use of statistics and its understanding and to foster its methodological developments.

- Benjamin Guedj has been named Young Leader of the Franco-British Council and received Knight of the Order of the Academic Palms of the French Republic.

- Sophie Dabo has been named Black Heroe of Mathematics of the Black Heroes of Mathematics 2022, at ICM, Edinburgh, 3-4, October 2022

7 New software and platforms

7.1 New software

7.1.1 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 Contributions: - 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

URL: https://github.com/modal-inria/MixtComp

Contact: Christophe Biernacki

Participants: Christophe Biernacki, Vincent Kubicki, Matthieu Marbac-Lourdelle, Serge Iovleff, Quentin Grimonprez, Etienne Goffinet

Partners: Université de Lille, CNRS

7.1.2 cfda

Name: Categorical functional data analysis

Keyword: Functional data

Functional Description: The R package cfda performs:

- descriptive statistics for categorical functional data

- dimension reduction and optimal encoding of states (correspondance multiple analyses towards functional data)

URL: https://github.com/modal-inria/cfda

Contact: Cristian Preda

Participants: Cristian Preda, Quentin Grimonprez, Vincent Vandewalle

Partner: Université de Lille

7.1.3 ClusPred

Name: Simultaneous Semi-Parametric Estimation of Clustering and Regression

Keywords: Regression, Clustering, Semi-parametric model, Finite mixture

Functional Description: Parameter estimation of regression models with fixed group effects, when the group variable is missing while group-related variables are available. Parametric and semi-parametric approaches are considered.

URL: https://cran.r-project.org/web/packages/ClusPred

Authors: Matthieu Marbac-Lourdelle, Mohammed Sedki, Christophe Biernacki, Vincent Vandewalle

Contact: Matthieu Marbac-Lourdelle

7.1.4 visCorVar

Name: visualization of correlated variables in the context of statistical integration of omics data

Keywords: Data integration, Visualization

Functional Description: The R package visCorVar allows visualizing results from data integration with the function block.spslda (bioconductor mixOmics package). The data integration is performed for different types of omic datasets (transcriptomics, metabolomics, metagenomics) in order to select variables of a omic dataset which are correlated with the variables of the other omic datasets and the response variables and to predict the class membership of a new sample. These correlated variables can be visualized with correlation circles and networks.

URL: https://gitlab.com/bilille/viscorvar

Contact: Guillemette Marot

Participants: Maxime Brunin, Guillemette Marot, Pierre Pericard

Partner: Université de Lille

7.1.5 metaRNASeq

Name: RNA-Seq data meta-analysis

Keywords: Transcriptomics, Meta-analysis, Differential analysis, High throughput sequencing, Biostatistics

Functional Description: MetaRNASeq is a specialised software for RNA-seq experiments. It is an R package which is an adaptation of the metaMA package, which performs meta-analysis of microarray data. Both enable to take advantage of empirical bayesian approaches, especially appropriate in a context of high dimension. Specificities of the two types of technologies require however some adaptations to each one, explaining the development of two different packages. To facilitate their use by a large public, a Galaxy-web instance named SMAGEXP has been created and gathers the two packages.

Release Contributions: Minimum maintenance was ensured to correct a bug reported by an user, due to Windows Systems, not appearing on Linux. This bug was related to the treatment of missing values. Guillemette Marot, who created and largely contributed to the initial versions of the metaRNASeq package, led the maintenance in September 2021 to Samuel Blanck, engineer in METRICS ULR2694 team (Univ. Lille, CHU Lille).

URL: https://cran.r-project.org/web/packages/metaRNASeq/index.html

Contact: Guillemette Marot

Participants: Guillemette Marot, Andrea Rau, Samuel Blanck

Partners: INRAE, Université de Lille

7.1.6 HDSpatialScan

Name: Multivariate and Functional Spatial Scan Statistics

Keywords: Functional data, Clustering, Spatial information, Multivariate data

Functional Description: Allows to detect spatial clusters of abnormal values on multivariate or functional data

URL: https://cran.r-project.org/web/packages/HDSpatialScan/index.html

Contact: Sophie Dabo

7.1.7 MLGL

Name: Multi-Layer Group Lasso

Keywords: Variable selection, Statistical learning

Functional Description: The MLGL R-package, standing for Multi-Layer Group-Lasso, implements a procedure of variable selection in the context of redundancy between explanatory variables, which holds true with high dimensional data. The MLGL approach combines variables aggregation and selection in order to improve interpretability and performance. First, a hierarchical clustering procedure provides at each level a partition of the variables into groups. Then, the set of groups of variables from the different levels of the hierarchy is given as input to group-Lasso, with weights adapted to the structure of the hierarchy. At this step, group-Lasso outputs sets of candidate groups of variables for each value of regularization parameter. The versatility offered by MLGL to choose groups at different levels of the hierarchy a priori induces a high computational complexity. MLGL however exploits the structure of the hierarchy and the weights used in group-Lasso to greatly reduce the final time cost. The final choice of the regularization parameter – and therefore the final choice of groups – is made by a multiple hierarchical testing procedure.

URL: https://cran.r-project.org/web/packages/MLGL/index.html

Contact: Guillemette Marot

7.2 New platforms

7.2.1 MASSICCC Platform

Participants: Christophe Biernacki, Julien Vandeale.

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. From 2020, Julien Vandaele joined the MODAL team as a research engineer for upgrading the MixtComp software and also for replacing the MASSICCC platform by some three R notebooks dedicated to the three packages Mixmod, BlockCluster and MixtComp. All these notebooks can be founded here on the MODAL webpage.

8 New results

8.1 Axis 1: Co-clustering as a (very) parsimonious clustering

Participants: Christophe Biernacki.

We advocate that co-clustering, is of particular interest to perform high dimension (HD) clustering of individuals even if it is not its primary mission. Indeed, column clustering is recast as a strategy to control the variance of the estimation, the model dimension being driven by the number of groups of variables instead of the number of variables itself. A survey paper [40] advocates the ability of co-clustering to outperform simple mixture row-clustering, even if co-clustering clearly corresponds to a misspecified model situation, revealing a promising manner to efficiently address (very) HD clustering.

8.2 Axis 1: Relaxing the identically distributed assumption in Gaussian co-clustering for high dimensional data

Participants: 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 [13], although allowing more flexibility, still maintains the very high degree of parsimony achieved by traditional co-clustering, thus allowing it to be still used in the HD context.

8.3 Axis 1: Dealing with Missing Data in Model-based Clustering through a MNAR Model

Participants: Christophe Biernacki.

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. Currently, a preprint is being finalized for submission to an international journal. Some talks on this topic and also on a more general topic on missing data and its impact on mixtures and clustering have been given this year in some conferences or workshops [25, 26].

It is a joint work with Claire Boyer from Sorbonne Université, Gilles Celeux from Inria Saclay, Julie Josse from Inria Montpellier, Fabien Laporte from Institut Pasteur and Matthieu Marbac from ENSAI.

8.4 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 etc. 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. A paper has now been accepted in an international journal [18].

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

8.5 Axis 1: A Binned Technique for Scalable Model-based Clustering on Huge Datasets

Participants: Filippo Antonazzo, Christophe Biernacki.

Clustering is impacted by the regular increase of sample sizes which provides opportunity to reveal information previously out of scope. However, the volume of data leads to some issues related to the need of many computational resources and also to high energy consumption. Resorting to binned data depending on an adaptive grid is expected to give proper answer to such green computing issues while not harming the quality of the related estimation. After a brief review of existing methods, a first application in the context of univariate model-based clustering is provided, with a numerical illustration of its advantages. Finally, an initial formalization of the multivariate extension is done, highlighting both issues and possible strategies. This work has been submited to an international journal (now under revision), and a PhD thesis on this topic has been defended [34].

It is a joint work with Christine Keribin from Université Paris-Saclay.

8.6 Axis 1: Forecasting elections results via the voter model with stubborn nodes

Participants: Benjamin Guedj, Antoine Vendeville.

We propose a novel method to forecast the result of elections using only official results of previous ones. It is based on the voter model with stubborn nodes and uses theoretical results developed in a previous work of ours. We look at popular vote shares for the Conservative and Labour parties in the UK and the Republican and Democrat parties in the US. We are able to perform time-evolving estimates of the model parameters and use these to forecast the vote shares for each party in any election. We obtain a mean absolute error of 4.74%. As a side product, our parameters estimates provide meaningful insight on the political landscape, informing us on the proportion of voters that are strong supporters of each of the considered parties.

8.7 Axis 1: Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly

Participants: Benjamin Guedj.

When confronted with massive data streams, summarizing data with dimension reduction methods such as PCA raises theoretical and algorithmic pitfalls. A principal curve acts as a nonlinear generalization of PCA, and the present work 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.

8.8 Axis 1&2: An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees

Participants: Christophe Biernacki, Guillaume Braun, Hemant Tyagi.

Real-world networks often come with side information that can help to improve the performance of network analysis tasks such as clustering. Despite a large number of empirical and theoretical studies conducted on network clustering methods during the past decade, the added value of side information and the methods used to incorporate it optimally in clustering algorithms are relatively less understood. We propose a new iterative algorithm to cluster networks with side information for nodes (in the form of covariates) and show that our algorithm is optimal under the Contextual Symmetric Stochastic Block Model. Our algorithm can be applied to general Contextual Stochastic Block Models and avoids hyperparameter tuning in contrast to previously proposed methods. We confirm our theoretical results on synthetic data experiments where our algorithm significantly outperforms other methods, and show that it can also be applied to signed graphs. Finally we demonstrate the practical interest of our method on real data.

This work has appeared in the proceedings of an international conference (ICML) [22].

8.9 Axis 1&2: Seeded graph matching for the correlated Wigner model via the projected power method

Participants: Ernesto Araya, Guillaume Braun, Hemant Tyagi.

The problem of matching two graphs (with n vertices) has many applications (*e.g.*, computer vision, network deanonymization *etc.*). We study this problem for weighted graphs under the setting where a partial match (also called seed) is provided [38]. Assuming the graphs are generated by the correlated Wigner model, we prove that the projected power method exactly recovers the latent matching after $O(\log n)$ iterations. Experiment results are provided on synthetic and real data to support our theoretical results. This work is currently under review in a journal.

8.10 Axis 1&2: Dynamic Ranking and Translation Synchronization

Participants: Ernesto Araya, Eglantine Karle, Hemant Tyagi.

In many applications, such as sport tournaments or recommendation systems, we have at our disposal data consisting of pairwise comparisons between a set of n items (or players). The objective is to use this data to infer the latent strength of each item and/or their ranking. Existing results for this problem predominantly focus on the setting consisting of a single comparison graph G. However, there exist scenarios (e.g., sports tournaments) where the pairwise comparison data evolves with time. Theoretical results for this dynamic setting are relatively limited and is the focus of this paper. We study an extension of the *translation synchronization* problem to the dynamic setting where the outcomes evolve smoothly over time, and derive efficient algorithms which are consistent (under a dynamic generative model) in terms of the number of time points. Experiments on synthetic and real data showcase the efficacy of the proposed methods. This work is currently under review in a journal [39].

8.11 Axis 1&2: Minimax Optimal Clustering of Bipartite Graphs with a Generalized Power Method

Participants: Guillaume Braun, Hemant Tyagi.

Clustering bipartite graphs is a fundamental task in network analysis, especially when the number of rows and columns of the adjacency matrix are of different order. Recent results provide an upper-bound for the misclustering rate when the columns (resp. rows) can be partitioned into $L = 2$ (resp. $K = 2$) communities. In this work [41] we introduce a new algorithm based on the power method and derive conditions for exact recovery in the general setting where $K = L \geq 2$. We also derive a minimax lower bound on the misclustering error when $K = L = 2$, which matches the upper bound up to a constant factor. This work is currently under review in a journal.

8.12 Axis 2: Asymptotic efficiency of some nonparametric tests for location on hyperspheres

Participants: Sophie Dabo-Niang.

In the paper, we show that several classical nonparametric tests for multivariate location in the Euclidean case can be adapted to nonparametric tests for the location problem on hyperspheres. The tests we consider are spatial signs and spatial signed-rank tests for location on hyperspheres. We compute the asymptotic powers of the latter tests in the classical rotationally symmetric case. In particular, we show that the spatial signed-rank based test uniformly dominates the spatial sign test and has performances that are extremely close to the asymptotically optimal test in the well-known von Mises-Fisher case. Monte-Carlo simulations confirm our asymptotic results.

It is a joint work with Baba Thiam (University of Lille, Painlevé), Thomas Verdebout (ULB, Belgium). This work has been been submitted for publication [43].

8.13 Axis 2: k-nearest neighbors prediction and classification for spatial data

Participants: Sophie Dabo-Niang.

This paper proposes a spatial k-nearest neighbor method for nonparametric prediction of real-valued spatial data and supervised classification for categorical spatial data. The proposed method is based on a double nearest neighbor rule which combines two kernels to control the distances between observations and locations. It uses a random bandwidth in order to more appropriately fit the distributions of the covariates. The almost complete convergence with rate of the proposed predictor is established red and the almost sure convergence of the supervised classification rule was deduced. Finite sample properties are given for two applications of the k-nearest neighbor prediction and classification rule.

It is a joint work with Mohamed Salem Ahmed (University of Lille, CERIM), Mohamed Attouch (University Sidi Bel Abbes, Algeria), Mamadou Ndiaye (UCAD, Senegal). This work is under revision [37].

8.14 Axis 2: Progress in Self-Certified Neural Networks

Participants: Benjamin Guedj.

A learning method is self-certified if it uses all available data to simultaneously learn a predictor and certify its quality with a tight statistical certificate that is valid with high confidence on any random data point. Self-certified learning promises to bring two major advantages to the machine learning community: First, it avoids the need to hold out data for validation and test purposes, both for certifying the model's performance as well as for model selection. This could lead to a simplification of the machine learning data pipeline, while additionally, using all the available data for training could also lead to better representations of the underlying data distribution and ultimately lead to more accurate models. Secondly, self-certified learning focuses on delivering performance certificates that are valid with high confidence and are informative of the out-of-sample error, properties that are crucial for appropriately comparing machine learning models as well as setting performance standards for algorithmic governance of these models in the real world. In this paper, we assess how close we are to achieving self-certification in neural networks. In particular, recent work has shown that probabilistic neural networks trained by optimising PAC-Bayes generalisation bounds could bear promise towards achieving self-certified learning, since these can leverage all the available data to learn a posterior and simultaneously certify its risk with tight statistical performance certificates. In this work we empirically compare (on 4 classification datasets) test set generalisation bounds for deterministic predictors and a PAC-Bayes bound for randomised predictors obtained by a self-certified learning strategy (i.e. using all available data for training). We first show that both of these generalisation bounds are not too far from test set errors. We then show that in data small regimes, holding out data for the test set bounds adversely affects generalisation performance, while self-certified strategies based on PAC-Bayes bounds do not suffer from this drawback, showing that they might be a suitable choice for this small data regime. We also find that self-certified probabilistic neural networks learnt by PAC-Bayes inspired objectives lead to certificates that can be surprisingly competitive compared to commonly used test set bounds.

8.15 Axis 2: MMD Aggregated Two-Sample Test

Participants: Benjamin Guedj, Antonin Schrab.

We propose a novel nonparametric two-sample test based on the Maximum Mean Discrepancy (MMD), which is constructed by aggregating tests with different kernel bandwidths. This aggregation procedure, called MMDAgg, ensures that test power is maximised over the collection of kernels used, without requiring held-out data for kernel selection (which results in a loss of test power), or arbitrary kernel choices such as the median heuristic. We work in the non-asymptotic framework, and prove that our aggregated test is minimax adaptive over Sobolev balls. Our guarantees are not restricted to a specific kernel, but hold for any product of one-dimensional translation invariant characteristic kernels which are absolutely and square integrable. Moreover, our results apply for popular numerical procedures to determine the test threshold, namely permutations and the wild bootstrap. Through numerical experiments on both synthetic and real-world datasets, we demonstrate that MMDAgg outperforms alternative state-of-the-art approaches to MMD kernel adaptation for two-sample testing.

8.16 Axis 2: Learning PAC-Bayes Priors for Probabilistic Neural Networks

Participants: Benjamin Guedj.

Recent works have investigated deep learning models trained by optimising PAC-Bayes bounds, with priors that are learnt on subsets of the data. This combination has been shown to lead not only to accurate classifiers, but also to remarkably tight risk certificates, bearing promise towards self-certified learning (i.e. use all the data to learn a predictor and certify its quality). In this work, we empirically investigate the role of the prior. We experiment on 6 datasets with different strategies and amounts of data to learn data-dependent PAC-Bayes priors, and we compare them in terms of their effect on test performance of the learnt predictors and tightness of their risk certificate. We ask what is the optimal amount of data which should be allocated for building the prior and show that the optimum may be dataset dependent. We demonstrate that using a small percentage of the prior-building data for validation of the prior leads to promising results. We include a comparison of underparameterised and overparameterised models, along with an empirical study of different training objectives and regularisation strategies to learn the prior distribution.

8.17 Axis 2: On Margins and Derandomisation in PAC-Bayes

Participants: Benjamin Guedj, Felix Biggs.

We give a general recipe for derandomising PAC-Bayesian bounds using margins, with the critical ingredient being that our randomised predictions concentrate around some value. The tools we develop straightforwardly lead to margin bounds for various classifiers, including linear prediction – a class that includes boosting and the support vector machine – single-hidden-layer neural networks with an unusual erf activation function, and deep ReLU networks. Further, we extend to partially-derandomised predictors where only some of the randomness is removed, letting us extend bounds to cases where the concentration properties of our predictors are otherwise poor. For more details see [20].

8.18 Axis 2: Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound

Participants: Benjamin Guedj, Valentina Zantedeschi.

We investigate a stochastic counterpart of majority votes over finite ensembles of classifiers, and study its generalization properties. While our approach holds for arbitrary distributions, we instantiate it with Dirichlet distributions: this allows for a closed-form and differentiable expression for the expected risk, which then turns the generalization bound into a tractable training objective. The resulting stochastic majority vote learning algorithm achieves state-of-the-art accuracy and benefits from (non-vacuous) tight generalization bounds, in a series of numerical experiments when compared to competing algorithms which also minimize PAC-Bayes objectives – both with uninformed (data-independent) and informed (data-dependent) priors. Fore more details see [23].

8.19 Axis 2: Differentiable PAC–Bayes Objectives with Partially Aggregated Neural Networks

Participants: Benjamin Guedj, Felix Biggs.

We make two related contributions motivated by the challenge of training stochastic neural networks, particularly in a PAC–Bayesian setting: (1) we show how averaging over an ensemble of stochastic neural networks enables a new class of partially-aggregated estimators, proving that these lead to unbiased lower-variance output and gradient estimators; (2) we reformulate a PAC–Bayesian bound for signed-output networks to derive in combination with the above a directly optimisable, differentiable objective and a generalisation guarantee, without using a surrogate loss or loosening the bound. We show empirically that this leads to competitive generalisation guarantees and compares favourably to other methods for training such networks. Finally, we note that the above leads to a simpler PAC–Bayesian training scheme for sign-activation networks than previous work. For more details see [21].

8.20 Axis 2: PAC-Bayes Unleashed: Generalisation Bounds with Unbounded Losses

Participants: Benjamin Guedj, Maxime Haddouche.

We present new PAC-Bayesian generalisation bounds for learning problems with unbounded loss functions. This extends the relevance and applicability of the PAC-Bayes learning framework, where most of the existing literature focuses on supervised learning problems with a bounded loss function (typically assumed to take values in the interval [0;1]). In order to relax this classical assumption, we propose to allow the range of the loss to depend on each predictor. This relaxation is captured by our new notion of HYPothesis-dependent rangE (HYPE). Based on this, we derive a novel PAC-Bayesian generalisation bound for unbounded loss functions, and we instantiate it on a linear regression problem. To make our theory usable by the largest audience possible, we include discussions on actual computation, practicality and limitations of our assumptions.

8.21 Axis 2: Still No Free Lunches: The Price to Pay for Tighter PAC-Bayes Bounds

Participants: 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, which is more or less realistic for a given problem. Some models are “expensive” (strong assumptions, such as sub-Gaussian 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 of assumptions minimal. The present paper explores and exhibits what the limits are for obtaining tight probably approximately correct (PAC)-Bayes bounds in a robust setting for cheap models.

8.22 Axis 3: Non-parametric statistical analysis of spatially distributed functional data

Participants: Sophie Dabo-Niang.

A nonparametric estimator of the regression function of a scalar spatial variable given a functional spatial variable is proposed. Mean square and almost complete consistencies of the estimator are obtained when the sample considered is an α-mixing sequence and composed of non i.i.d observations. Lastly, an application to spatial prediction and numerical results are provided to illustrate the behavior of our estimator.

It is a joint work with Baba Thiam (University of Lille), Camille Ternynck (University of Lille, CERIM), Anne-Françoise Yao (University of Clermont Auvergne).

8.23 Axis 3: Clustering spatial functional data

Participants: Vincent Vandewalle, Cristian Preda, Sophie Dabo-Niang.

In this work we present two approaches for clustering spatial functional data. The first one is the model-based clustering that uses the concept of density for functional random variables. The second one is the hierarchical clustering based on univariate statistics for functional data such as the functional mode or the functional mean. These two approaches take into account the spatial features of the data: two observations that are spatially close share a common distribution of the associated random variables. The two methodologies are illustrated by an application to air quality data.

8.24 Axis 3: Regression models for spatially distributed autoregressive functional data

Participants: Sophie Dabo-Niang.

A functional linear autoregressive spatial model, where the explanatory variable takes values in a function space while the response process is real-valued and spatially autocorrelated, is proposed. The specificity of the model is due to the functional nature of the explanatory variable and the structure of a spatial weight matrix that defines the spatial dependency between neighbors. The estimation procedure consists of reducing the infinite dimension of the functional explanatory variable and maximizing the quasi-maximum likelihood. We establish the consistency and asymptotic normality of the estimator. The ability of the methodology is illustrated via simulations and by application to real data.

It is a joint work with Mohamed Salem Ahmed (University of Lille, CERIM), Zied Gharbi (University of Lille) Laurence Broze (Unievrsity of Lille).

8.25 Axis 3: Investigating spatial scan statistics for multivariate functional data

Participants: Sophie Dabo-Niang.

This work presents the R package HDSpatialSca [44] that allows users to apply easily spatial scan statistics on real-valued multivariate data or both univariate and multivariate functional data. It also permits to plot the detected clusters and to summarize them. In this article the methods are presented and the use of the package is illustrated through examples on environmental data provided in the package.

It is a joint work with Camille Frévent (University of Lille, CERIM), Mohamed-Salem Ahmed (University of Lille, CERIM), Michaël Genin (University of Lille, CERIM).

8.26 Axis 3: PLS regression approach for multivariate functional data with different domains

Participants: Cristian Preda, Issam Moindjie, Sophie Dabo.

Multivariate functional data is considered as sample paths of a multivariate valued stochastic process, $X = (X_1, \dots, X_d)$. In this setting, each dimension X_i, $i = 1, \dots, d$, is a stochastic process, $X_i = \{X_i(t), t \in \mathscr{I}_i\}$, where $\mathscr{I}_i$ is some compact domaine of $\mathbb{R}$. The problems of linear regression and binary classification are addressed by PLS regularization techniques. For application purposes, decision tree methods combined with functional PLS regression are proposed [47].

8.27 Axis 3: Group lasso regression for spatially dependent functional data

Participants: Cristian Preda, Issam Moindjie, Sophie Dabo.

Multivariate functional data is considered under the assumption of spatially dependence between dimensions. Each dimension is associated to some (spatial) clusters with potentially different effect on a response variable. In the context of linear regression with multivariate functional data, a natural assumption is to consider the same regression coefficient (slope) function for all dimensions belonging to the same cluster. Fused and group lasso techniques are extended for this purpose.

8.28 Axis 4: Statistical analysis of high-throughput proteomic data

Participants: Guillemette Marot, Vincent Vandewalle, Wilfried Heyse.

Since November 2019, Wilfried Heyse has started a PhD thesis granted by INSERM and supervised by Christophe Bauters, Guillemette Marot and Vincent Vandewalle. The aim is to identify earlier after myocardial infarction (MI) patients at high risk of developing left ventricular remodelling (LVR) that is quantified by imaging one year after MI or to identify patients with high risk of death. For that purpose, high throughput proteomic approach is 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 have been collected on each of the 200 patients. In 2022, we have improved the selection of the proteomic signature by taking into account competing risks to answer clinical and statistical concerns of the reviewers and submitted a new version of the main paper of the PhD thesis. We also presented related work on score building in an invited seminar [30]. In parallel, we have studied complementary measures (repeated measurements on 4 time points). The aim of our current work is to jointly model the temporal structure of all the proteins and the long-term survival of the patients. This is a joint work with Florence Pinet from INSERM.

8.29 Axis 4: Multi-layer group Lasso

Participants: Quentin Grimonprez, Guillemette Marot.

Multi-Layer Group-Lasso (MLGL) is a procedure of variable selection in the context of redundancy between explanatory variables, which holds true with high-dimensional data. The proposed approach combines variable aggregation and selection in order to improve interpretability and performance. The associated R package is available on CRAN and the associated publication has been presented orally at useR! 2022 conference [29]. The publication [15] gives more details about the statistical procedure.

8.30 Axis 4: Statistical analysis of transcriptomics data

Participants: Guillemette Marot.

Thanks to multivariate sparse Partial Least Squares-Discriminant Analysis, we selected eight miRNAs which discriminated severe and non severe COVID-19 patients. We discussed the interest of these miRNAs for further score building in an accepted publication [14]. We were also invited in a workshop to present past results on meta-analysis of RNA-Seq data [28]. Questions of participants gave us new ideas for further maintenance of the metaRNASeq package, which will be performed in 2023.

8.31 Axis 4: Statistical analysis of proteomic data with empirical bayesian approaches

Participants: Guillemette Marot.

Our expertise on empirical bayesian approaches for proteomics data analysis has led to a publication in a biological journal with impact factor 28: Annals of Rheumatic Diseases [51]. This is a joint work with Dr S.Sanges and Pr D. Launay. The proteomic analysis has revealed potential biomarkers that may assist diagnosis and treatment of patients with systemic clerosis-associated pulmonary arterial hypertension (SSc-PAH). Further biological validation in an independent cohort revealed that chemerin, which was highlighted in the exploratory analysis, was a reliable surrogate biomarker for pulmonary vascular resistance.

8.32 Axis 4: Multi-omics data analysis

Participants: Guillemette Marot.

Using sparse generalized canonical correlation analysis in addition to standard omics analyses, we have performed several multi-omics analyses, one being published in [11]. This paper is a joint work with A. Chepy, Pr. D. Launay, Pr. V. Sobanski and members from platforms generating or analysing transcriptomics and proteomics data. It explores the role of purified immunoglobulins from patients suffering from systemic sclerosis on phenotypes of fibroblasts.

8.33 Axis 4: Interpretable Domain Adaptation for Hidden Subdomain Alignment in the Context of Pre-trained Source Models

Participants: Christophe Biernacki, Luxin Zhang.

Domain adaptation aims to leverage source domain knowledge to predict target domain labels. Most domain adaptation methods tackle a single-source, single-target scenario, whereas source and target domain data can often be subdivided into data from different distributions in real-life applications (e.g., when the distribution of the collected data changes with time). However, such subdomains are rarely given and should be discovered automatically. To this end, some recent domain adaptation works seek separations of hidden subdomains, w.r.t. a known or fixed number of subdomains. In contrast, this paper introduces a new subdomain combination method that leverages a variable number of subdomains. Precisely, we propose to use an inter-subdomain divergence maximization criterion to exploit hidden subdomains. Besides, our proposition stands in a target-to-source domain adaptation scenario, where one exploits a pre-trained source model as a black box; thus, the proposed method is model-agnostic. By providing interpretability at two complementary levels (transformation and subdomain levels), our method can also be easily interpreted by practitioners with or without machine learning backgrounds. Experimental results over two fraud detection datasets demonstrate the efficiency of our method. This work has been accepted to an international conference [33].

It is a joint work with Pascal Germain from Université Laval (Canada) and with Yacine Kessaci from Worldline company.

8.34 Axis 4: Interpretable Domain Adaptation Using Unsupervised Feature Selection on Pretrained Source Models

Participants: Christophe Biernacki, Luxin Zhang.

We study a realistic domain adaptation setting where one has access to an already existing “black-box” machine learning model. Indeed, in real-life scenarios, an efficient pre-trained source domain predictive model is often available and required to be preserved. The solution we propose to this problem has the asset to provide an interpretable target to source transformation, by seeking a sparse and ordered coordinate-wise adaptation of the feature space, in addition to elementary mapping functions. To automatically select the subset of features to be adapted, we first introduce a weakly-supervised process relying on scarce labeled target data. Then, we address a more challenging unsupervised version of this domain adaptation scenario. To this end, we propose a new pseudo-label estimator over unlabeled target examples, which is based on the rank-stability in regards to the source model prediction. Such estimated “labels” are further used in a feature selection process to assess whether each feature needs to be transformed to achieve adaptation. We provide theoretical foundations of our method as well as an efficient implementation. Numerical experiments on real datasets show particularly encouraging results since approaching the supervised case, where one has access to labeled target samples. This work has been now published in an international journal [19].

It is a joint work with Pascal Germain from Université Laval (Canada) and with Yacine Kessaci from Worldline company.

8.35 Axis 4: Single cell classification using statistical learning on mechanical properties measured by mems tweezers

Participants: Sophie Dabo-Niang.

Cell population is heterogenous and so presents a wide range of properties as metastatic potential. But using rare cells for clinical applications requires precise classification of individual cells. Here, we propose a multi-parameter analysis of single cells to classify them using statistical learning techniques and to predict the sub-population of each cell, although they may have close characteristics. We used MEMS tweezers to analyze mechanical properties (stiffness, viscosity, and size) of single cells from two different breast cancer cell lines in a controlled environment and run supervised learning methods to predict the population they belong to. This label-free method is a significant step forward to distinguish rare cell sub-populations for clinical applications.

This work has been presented to the international conference "The 35th International Conference on Micro Electro Mechanical Systems", on January 2022 [50].

It is a joint work with Dominique Collard (LIMMS, CNRS, Universities of Lille and Tokyo), Cagatay Mehmed (LIMMS, CNRS, Universities of Lille and Tokyo) and others colleagues from University of Tokyo.

8.36 Axis 4: Dimensionality Reduction and Bandwidth Selection for Spatial Kernel Discriminant Analysis

Participants: Sophie Dabo-Niang.

Spatial Kernel Discriminant Analysis is a powerful tool for the classification of spatially dependent data. It allows taking into consideration the spatial autocorrelation of data based on a spatial kernel density estimator. The performance of SKDA is highly influenced by the choice of the smoothing parameters, also known as bandwidths. Moreover, computing a kernel density estimate is computationally intensive for high-dimensional datasets. In this paper, we consider the bandwidth selection as an optimization problem, that we resolve using Particle Swarm Optimization algorithm. In addition, we investigate the use of Principle Component Analysis as a feature extraction technique to reduce computational complexity and overcome curse of dimensionality drawback. We examined the performance of our model on Hyperspectral image classification. Experiments have given promising results on a commonly used dataset.

This work has been presented in the 13th International Conference on Agents and Artificial Intelligence ICAART and published in the proceeding.

It is a joint work with Soumia Boumeddane, Leila Hamdad, Hamid Haddadou (ESI, Algeria).

8.37 Axis 4: A kernel discriminant analysis for spatially dependent data

Participants: Sophie Dabo-Niang.

We propose a novel supervised classification algorithm for spatially dependent data, built as an extension of kernel discriminant analysis, that we named Spatial Kernel Discriminant Analysis (SKDA). Our algorithm is based on a kernel estimate of the spatial probability density function, which integrates a second kernel to take into account spatial dependency of data. In fact, classical data mining algorithms assume that data samples are independent and identically distributed. However, this assumption is not verified when dealing with spatial data characterized by spatial autocorrelation phenomenon. To make an accurate analysis, it is necessary to exploit this rich source of information and to capture this property. We have applied our algorithm to a relevant domain, which consist of the classification of remotely sensed hyperspectral images. In order to assess the efficiency of our proposed method, we conducted experiments on two remotely sensed images datasets (Indian Pines and Pavia University) with different characteristics and scenarios. The experimental results show that our method is competitive and achieves higher classification accuracy compared to other contextual classification methods.

This work has been published in Distributed and Parallel Databases. It is a joint work with Soumia Boumeddane, Leila Hamdad, Hamid Haddadou (ESI, Algeria).

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

Diagrams Technologies startup

Participants: Christophe Biernacki, Cristian Preda.

Christophe Biernacki and Cristian Preda act as scientific experts for the Diagrams Technologies startup specialized in industrial data analysis a software dedicated to predictive maintenance. This startup is a spinoff of the MODAL team.

Program France-Relance : MODAL-Alicante

Participants: Cristian Preda, Vincent Vandewalle.

The objective of this collaboration is to develop statistical learning models that explore the temporal dimension of health data within the framework of projects developed by the company ALICANTE and whose solutions are provided by the research work of the MODAL team. Ismat Draa and Rachid Boulkhir are part of this project.

Duration: 2 years (15/12/2021 - 15/12/2023)

ADULM

Participants: Sophie Dabo-Niang, Cristian Preda.

The main goal of this projet with Lille Metropole Urban Development and Planning Agency (ADULM) is to design a tool for Territorial Coherence Scheme (SCoT) to monitor urban developments and develop territorial observation

Saint-Gobain

Participants: Christophe Biernacki, Vincent Vandewalle, Myriam Benbahlouli.

Saint-Gobain designs, produces and distributes materials and solutions for the construction, mobility, healthcare and other industrial applications. The purpose of this contract is to perform multi-product forecast. This work has been initiated during the internship of Myriam Benbahlouli at Saint-Gobain during July and August. This work continues with Myriam Benbahlouli's apprentice contract at Inria. This work in done under the supervision of Christophe Biernacki and Vincent Vandewalle.

9.2 Bilateral grants with industry

Worldline

Participants: Christophe Biernacki, Alain Celisse.

Worldline is the new world-class leader in the payments and transactional services industry, with a global reach. A PhD began in Feb. 2019 with Luxing Gang under the supervision of Christophe Biernacki, Pascal Germain (Laval University, Canada) and Yacine Kessaci (Worldline) on the topic of the domain adaptation from a pre-trained source model (with application to fraud detection in electronic payments). A second Phd thesis with Etienne Kronert started in September 2020. This thesis issupervised by Alain Celisse and it aims to develop techniques for anomaly detection in financial time series.

ADEO

Participants: Christophe Biernacki, Vincent Vandewalle.

Adeo is No. 1 in Europe and No. 3 worldwide in the DIY market. A PhD began in Dec. 2020 with Axel Potier under the supervision of Christophe Biernacki, Vincent Vandewalle, Matthieu Marbac (ENSAI) and Julien Favre (ADEO) on the topic of sales forecasting concerning "slow movers" items (equivalent to item sold in low quantities).

Seckiot

Participants: Christophe Biernacki, Cristian Preda.

Seckiot is an editor of cybersecurity software to protect industrial systems & IoT. From December 2021, Clarisse Boinay begun her Cifre PhD thesis (with AID, Agence de l'Innovation de Défense) with Seckiot on the topic of "anomaly detection and change point detection in contextual dynamic asynchronous graphs with applications in OT cybersecurity" under the co-supervision of Thomas Anglade (Seckiot), Christophe Biernacki and Cristian Preda.

Decathlon

Participants: Cristian Preda.

Decathlon is a brand specializing in the large distribution of sports equipment and materials. From September 2022, François Bassac begun his PhD thesis within Inria-Decathlon partnership on the topic of predicting performances and injuries with training data under the supervision Cristian Preda.

ASYGN

Participants: Sophie Dabo, Cristian Preda, Vincent Vandewalle.

ASYGN is a company specialized on the signal treatment chain. Modal is working with this compagny and LIMMS/CNRS-IIS to apply bioMEMS technology in the field of cancer

HORIBA

Participants: Sophie Dabo, Cristian Preda, Vincent Vandewalle.

HORIBA is a company specialized on optical spectrometry. Modal is working with this compagny and CENTRAL Lille on Raman spectroscopy and Artificial Intelligence dedicated to the synthesis in chemistry

10 Partnerships and cooperations

10.1 International initiatives

Gendergap with IMU- International Mathematical Union

Participants: Sophie Dabo.

Gendergap (2021 - 2023) Sophie Dabo-Niang is PI of the Gendergap global survey of scientists: a focus on Africa and Mathematics; by IMU (International Mathematical Union)

10.1.1 Visits to international teams

Research stays abroad

Sophie Dabo

AIMS:

South Africa:

Dates: 16-23 July, 2022

Context of the visit: CIMPA School

Mobility program/type of mobility: (supervision of a research school)

AIMS:

Senegal:

Dates: 30 March-4 April ,2022

Context of the visit: ARTS School

Mobility program/type of mobility: (organization of a research school)

ICM:

UK, Edinburgh:

Dates: -16 September, 2022

Context of the visit: EMS (European-Mathematical-Society)

Mobility program/type of mobility: (chair of CDC and presidents annual meeting)

10.2 European initiatives

10.2.1 H2020 projects

H2020 FAIR

Participants: Guillemette Marot.

- Acronym: FAIR
- Project title: Flagellin aerosol therapy as an immunomodulatory adjunct to the antibiotic treatment of drug-resistant bacterial pneunomia
- Coordinator: JC Sirard
- Duration: 4 years (2020-2023)
- Partners: Inserm (France), Univ Lille (France), Freie Universitaet Berlin (DE), Epithelix (CH), Aerogen (IE), Statens Serum Institut (DK), CHRU Tours (France), Academisch Medisch Centrum bij de Universiteit van Amsterdam (NL), University of Southampton (UK), European respiratory society (CH)
- Abstract: FAIR, project coordinated by JC. Sirard (Inserm, CIIL), aims at evaluating an alternative adjunct strategy to standard of care antibiotics for treating pneumonia caused by antibiotic-resistant bacteria: activation of the innate immune system in the airways. Guillemette Marot is involved in this H2020 project as scientific head of bilille platform, and will supervise 1 year engineer on analysis of omics data.

10.3 National initiatives

"Inria Challenge" ROAD-AI with Cerema

Participants: Vincent Vandewalle, Christophe Biernacki, Cristian Preda.

Cerema (Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement - Centre for Studies on Risks, the Environment, Mobility and Urban Planning) is a public institution dedicated to supporting public policies, under the dual supervision of the ministry for ecological transition and the ministry for regional cohesion and local authority relations. MODAL is involved in the ROAD-AI (Routes et Ouvrages d'Art Diversiformes, Augmentés & Intégrés) "Inria Challenge", with five other Inria teams (ACENTAURI, COATI, FUN, STATIFY, TITANE) including statistics, robotics, telecomunication, sensors network and 3D modeling. This four year project (starting in 2021) aims at having more sustainable, safer and more resilient transport infrastructures. It led to a general presentation to a conference [24] dedicated to experts on transport infrastructure management.

Program "Action Exploratoire" PATH : METRICS and CHU Lille

Participants: Sophie Dabo (coordinator), Vincent Vandewalle, Christophe Biernacki, Guillemette Marot, Cristian Preda.

The research project is part of an INRIA exploratory action by a consortium of doctors, bio-statisticians and statisticians. The aim is to provide a better understanding of the key stages in the patient's care pathway by bringing together the producers of data as close to the patient as possible, those who manage them, those who pre-process them, and those who analyse them, in order to obtain results as close to the field as possible and to provide the most efficient feedback to the clinician and the patient.

The project, which is essentially interdisciplinary and exploratory, is a continuation of past collaborations between members of the two units INRIA-MODAL and METRICS (University of Lille/CHU Lille). It could not be carried out without close collaboration between doctors and researchers in applied mathematics.

The analysis of care pathways and their adequacy to needs and resources has thus become a major scientific and administrative challenge. Although the digital data available for this purpose is increasing rapidly, the statistical methods and tools available to researchers and health authorities remain limited and inefficient.

The types of care pathways are very numerous. As part of this exploratory action, we propose to focus on two cases of application: 1) an ambulatory care pathway (city-hospital link); 2) an intra-hospital care pathway. This choice is justified by METRICS' solid expertise in these pathways, based on several years of research, as well as close links with clinicians who are experts in these issues.

Duration: 2 years (1/09/2021 - 31/08/2023)

Industrial Chair Smart digicat

Participants: Vincent Vandewalle, Cristian Preda, Sophie Dabo.

SmartDigiCat is a project led by Sebastien Paul (Professor at Centrale Lille, researcher at Unité de Catalyse et Chimie du Solide (UCCS – UMR CNRS 8181)) and involving several companies (SOLVAY, HORIBA, TEAMCAT SOLUTIONS) and academic laboratories (UCCS, CRIStAL, Inria and l'Institut Eugène Chevreul).

The consortium of the SmartDigiCat chair will develop an innovative approach for safer and more environmentally-friendly catalytic processes design. The innovation will emerge from the powerful combination of high-throughput experiments, theoretical chemistry and artificial intelligence. The domains of application of the tools developed for catalysis will be extended, among others, to materials and formulations.

Vincent Vandewalle, Cristian Preda and Sophie Dabo are implicated in the artificial intelligence part of the project. This part requires functional data analysis tools and challenging developments, for example to optimize the chemical process in order to obtain a target spectrum.

French Institute of Bioinformatics and equipex+ MuDiS4LS

Participants: Guillemette Marot.

- **Coordinators:** Claudine Medigue and Jacques Van Helden (Co-heads IFB)
- **Duration:** 7 years (2021 – 2028)

- **Abstract:** Bilille, the bioinformatics platform of Lille, is a member of IFB, the French Institute of Bioinformatics. IFB has obtained the funding of equipex+ MuDiS4LS (Mutualised Digital Spaces for FAIR data in Life and Health Science). As the scientific head of bilille platform, Guillemette Marot is also the scientific head of Univ. Lille partner for this equipex+. As a researcher, she will participate to implementation studies involving integration of complex data (IS1 and IS4). More information given by IFB and on bilille website.

10.3.1 ANR

CYTOMEMS

Participants: Sophie Dabo, Cristian Preda, Vincent Vandewalle.

- **Type:** ANR AAPG
- **Acronym:** CYTOMEMS
- **Project title:** Smart MEMS Instrumentation for Biophysical flow Cytometry with Statistical Learning
- **Coordinator:** Dominique Collard (CNRS)
- **Duration:** 2022–2024
- **Funding:** 600k EUR
- **Partners:** MODAL, Laboratoire Hubert Curien (UMR LIMMS CNRS IMU 2820)

APRIORI

Participants: Benjamin Guedj, Hemant Tyagi.

- **Type:** ANR PRC
- **Acronym:** APRIORI
- **Project title:** PAC-Bayesian theory and algorithms for deep learning and representation learning
- **Coordinator:** Emilie Morvant (Université Jean Monnet)
- **Duration:** 2019–2023
- **Funding:** 300k EUR
- **Partners:** MODAL, Laboratoire Hubert Curien (UMR CNRS 5516)

BEAGLE

Participants: Benjamin Guedj *(coordinator)*, Pascal Germain.

- **Type:** ANR JCJC
- **Acronym:** BEAGLE
- **Duration:** 2019–2023
- **Project title:** PAC-Bayesian theory and algorithms for agnostic learning
- **Funding:** 180k EUR
- **Partners:** Pierre Alquier (RIKEN AIP, Japan), Peter Grünwald (CWI, The Netherlands), Rémi Bardenet (UMR CRIStAL 9189)

SMILE

Participants: Christophe Biernacki, Vincent Vandewalle.

- **Acronym:** SMILE
- **Duration:** 2018–2022
- **Project title:** Statistical Modeling and Inference for unsupervised Learning at LargE-Scale)
- **Coordinator:** Faicel Chamroukhi (LMNO, Université de Caen)
- **Partners:** MODAL, LMNO UMR CNRS 6139 (Caen), LMRS UMR CNRS 6085 (Rouen), LIS UMR CNRS 7020 (Toulon)

TransEAsome

Participants: Guillemette Marot.

- **Type:** AMI Maladies rares
- **Acronym:** TransEAsome
- **Duration:** 72 months (2022 - 2027)
- **Project title:** Long term outcome of esophageal atresia: transomics profiles in adolescence
- **Funding:** 1.4M euros
- **Coordinator:** F. Gottrand (Infinite)
- **Partners:** Univ. Lille, Inserm NO, Inserm ADR - GO, CRACMO, FIMATHO

Oesomics

Participants: Guillemette Marot.

- **Type:** ANR AAP Recherche translationnelle en santé
- **Acronym:** Oesomics
- **Duration:** 36 months (2022 - 2027)
- **Project title:** Molecular signatures of esophageal atresia: towards the identification of the molecular causes of the different forms of esophageal atresia and prenatal diagnosis
- **Funding:** 233k euros
- **Coordinator:** F. Gottrand (Infinite)
- **Partners:** CHU Lille, PRISM, PLBS-Goal, PLBS-bilille

10.3.2 RHU and FHU

A RHU (recherche hospitalo-universitaire) is an excellence programme funded by PIA (program of investment for the future) and selected by ANR. A FHU is a federative project and a label necessary to postulate for a RHU.

RHU PreciNASH

Participants: Guillemette Marot.

- **Acronym:** PreciNASH
- **Project title:** Non-alcoholic steato-hepatitis (NASH) from disease stratification to novel therapeutic approaches
- **Coordinator:** François Pattou (Université de Lille, CHU Lille)
- **Duration:** 6 years (2016 –2022)
- **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 has supervised a 2 years post-doc, as her team ULR 2694 METRICS is a member of the FHU Integra. She also has supervised during two years an engineer of bilille platform for this project. METRICS is involved in the WP1 for the development of a clinical-biological model for the prediction of NASH. Bilille is involved in the task which consists to better stratify patients using unsupervised clustering. 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. More information on this project at PreciNASH project.

FHU PRECISE

Participants: Guillemette Marot, Christophe Biernacki.

- **Coordinator:** Pr D. Launay (U. Lille, CHU Lille)
  - **Acronym:** PRECISE
  - **Project title:** PREcision health in Complex Immune-mediated inflammatory diseaSEs
  - **Duration:** 5 years (2021 – 2025)
  - **Partners:** CHU Lille, CHU Amiens, CHU Rouen, CHU Caen, Université de Lille, Université de Picardie, Université de Rouen, Inserm
  - **Abstract** The objective of FHU PRECISE is to structure care, research and teaching relative to care of patients who suffer from complex IMID (Immune mediated inflammatory diseases) with an interdisciplinary approach. Guillemette Marot is the co-head with Vincent Sobanski and Grégoire Ficheur of the WP2 workpackage, which aims at creating a « virtual patient » and cluster patients based on their clinical and omic profiles. In this WP, she is involved both in the analysis task with bilille platform and in the research task led by Christophe Biernacki, involving MODAL team. This research task aims at combining complex data and integrating temporal structure in order to identify patient's care pathways. Guillemette Marot is also participating with bilille in WP3 for the research of a molecular signature predictive of the treatment response (resistance and complication).

10.3.3 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
  - Franco-African IRN (International Research Network) in Mathematics, funded by CNRS
  - ONCOLille (Cancer Research Institute in Lille)
- 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

10.4 Regional initiatives

Collaborations of the year linked to bilille

Participants: Guillemette Marot.

Bilille, the bioinformatics platform of Lille, has offered opportunities of collaborations with teams in biology and Health for projects with local partners. Guillemette Marot has supervised the data analysis part for the following research projects involving engineers from bilille (only the names of the principal investigators of the project are given even if several partners are sometimes involved in the project):

- LilNCog, M.-C. Chartier-Harlin, ANR Synapark
- Infinite, L. Dubuquoy, DiagOH
- PLBS, R. Viard, DTP
- U1011, J. Dubois-Chevalier
- PRISM, T. Cardon
- Laboratoire de Virologie, I. Engelmann

ONCOLille

Participants: Sophie Dabo, Cristian Preda, Vincent Vandewalle, Guillemette Marot, Christophe Biernacki.

ONCOLille, the cancer research institute of Lille, a interdisciplinary research institute combining biology, physics, chemistry, mathematics, bioinformatics, economics, health technologies, and human and social sciences by developing strong basic research and translational/pre-clinical research (development of alternative and original study models) in order to move towards transfer to the clinic (clinical trials, new molecules). Sophie Dabo is head of the mathematical team.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

Participants: Sophie Dabo, Guillemette Marot, Cristian Preda.

Sophie Dabo-Niang Co-organizer of ONCOLille days, November 2-4, 2022, Lille. (link)

Co-organizer of *20th International Workshop on Spatial Econometrics and Statistics*, 19-20 May 2021, Lille, France. (link)

Co-organizer of "Statistics and Science for health", June, 2022, Lille. (link)

Guillemette Marot Co-organizer of a workshop about bioinformatics and biostatistics: Learning for omics data integration, Lille, March 2022.

Cristian Preda Member of the Organizing commitee of the THE 23rd CONFERENCE of the ROMANIAN SOCIETY of PROBABILITY and STATISTICS, November 18-19, 2022. (link)

11.1.2 Scientific events: selection

Member of the conference program committees

Sophie Dabo-Niang is member of the conference program committee of ONCOLille days, November 2-4, 2022, Lille; *20th International Workshop on Spatial Econometrics and Statistics*, 19-20 May 2021, Lille, France; "Statistics and Science for health", June, 2022, Lille; 70th Congress of the French Economic Association (AFSE), 14-16 May 2022.

Guillemette Marot was a member of the scientific committee for the workshop about Learning for omics data integration, Lille, March 2022

Reviewer

Sophie Dabo-Niang is reviewer of the conference program committee of ONCOLille days, November 2-4, 2022, Lille; *20th International Workshop on Spatial Econometrics and Statistics*, 19-20 May 2021, Lille, France; "Statistics and Science for health", June, 2022, Lille; 70th Congress of the French Economic Association (AFSE), 14-16 May 2022.

11.1.3 Journal

Member of the editorial boards

Christophe Biernacki From several years: Associate Editor of the North-Western European Journal of Mathematics (NWEJM).

Sophie Dabo-Niang Since 2015, Associate editor of *Revista Colombiana de Estadística*
Since 2020, Associate editor of *Journal of Statistical Modeling and Analytics*
Since 2021, Associate editor of *Journal of Nonparametric Statistics*

Cristian Preda Since 2015, Associate editor for the Journal *Methodology and Computing in Applied Probability*

Reviewer - reviewing activities

Christophe Biernacki acted as a reviewer for some journals and conferences in the statistical and the machine learning community (CSDA, STCO, JOC, JCGS, LSTA, CaP).

Sophie Dabo-Niang acted as a reviewer for several journals (Statistical Inference for Stochastic Processes, Computational Statistics and Data Analysis, Statistics, Journal Afrika Statistika, Journal of Multivariate Analysis, Journal of Nonparametric statistics, Annales de l'ISUP, Electronic journal of statistics, Metika, Annals of Statistics,...).

Cristian Preda acted as a reviewer for journals and conferences in the statistics (CSDA, MCAP, IWAP).

Hemant Tyagi acted as a reviewer for conferences (ICML, NeurIPS, ICLR, LOG) and journals (Transactions on Signal and Information Processing over Networks, ACHA, SIAGA, Mathematical reviews).

11.1.4 Invited talks

Christophe Biernacki gave an invited talk to the international CMStatistics 2022 conference [25] and to the ASA 2022 workshop [26]

Guillemette Marot gave several invited talks:

- Thematic school FTMS Fourier transform mass spectrometry: treatment of ultra high resolution data, Ambleteuse, France (link) 17 November 2022
- Workshop ICM/Inria/SCAI Computational and mathematical approaches for neuroscience 8 June 2022 [30]
- Thematic day "RNA-Seq" organized by bilille and genomic platform of Lille, 28 June 2022 [28]
- 2 French lab seminaries (I2BC, Gif sur Yvette, 13 July 2022 and SAMM, Paris, 11 March 2022)

Sophie Dabo has been invited to talks:

- Colloque ENSAE, IRD, University Paris Dauphine, November, 1, 2022
- CISEN, 13-15 May 2022, Tunisia
- Black Heroes of Mathematics ; ICM, Edinburgh, UK, October 4-5 October, 2022. (link)
- Girls and Sciences days Artificial Intelligence Applied to Health, 4 November 2022, IN-RIA, Lille
- PASRC conference (orgnaized by Princton University), Abuja, December 12-15, 2022.

11.1.5 Leadership within the scientific community

Christophe Biernacki was elected as a Vice-head of the SFdS (Société Française de Statistique) since June 2022, which is the French society specialized in statistics, whose mission is to promote the use of statistics and its understanding and to foster its methodological developments.

Sophie Dabo-Niang Chair of EMS-CDC: 2019-2022; Chair of the Mathematics group ONCOLILLE (multidisciplinary Cancer Research Center), 2019-. Co-Chair of the working group in Mathematics of DSI-NRF-CNRS Workshop on Research for Impact Strengthening scientific collaboration between Europe and Africa, October 2022.

Guillemette Marot is the scientific head of bilille platform, labelled by IBISA and member platform of the French Institute of Bioinformatics.

11.1.6 Scientific expertise

Christophe Biernacki reviewed one CIFRE industrial PhD project for the ANRT institution.

Sophie Dabo-Niang has been expert on L'OREAL-UNESCO "Womens in Science" Awards; Member of the evaluation committee of the IBNI Prize with the support of IMU, LMS (London Mathematical Society), SFdS, SMAI, SMF, CIMPA, AMU.

11.1.7 Research administration

Christophe Biernacki Since January 2020, deputy scientific director of Inria at the national level in charge of the domain “Applied mathematics, computation and simulation".

Sophie Dabo-Niang Chair of EMS-CDC: 2019-2022; Member of the director committee of ONCOLILLE (multidisciplinary Cancer Research Center), 2019-. Co-Chair of the working group in Mathematics of DSI-NRF-CNRS Workshop on Research for Impact Strengthening scientific collaboration between Europe and Africa, October 2022.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Hemant Tyagi is teaching
  - Master: Statistics I, 24h, M1, Centrale Lille, France
  - Master: Statistics II, 24h, M1, Centrale 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
- Guillemette Marot is teaching
  - Licence: Biostatistics, 16.5h, L1, Université de Lille (Faculty of Medicine), France
  - Master: Biostatistics, 57.5h, M1, Université de Lille (Faculty of Medicine), France
  - Master: Supervised classification, 20h, M1, Polytech’Lille, France
  - Master: Biostatistics, 54h, M1, Université de Lille (Departments of Computer Science and Biology), France
  - Master: Artificial intelligence and health, M2, 3h, Université de Lille (Graduate school precision Health), France
  - Master: Statistical analysis of omic data, 12h, M2, Université de Lille (Department of Mathematics), France
  - Doctorat: Introduction to statistical analysis of omic data, 11h, Université de Lille (Faculty of Medicine), France
  - Doctorat: Statistical analysis of RNA-Seq data, 8h, Université de Lille (Faculty of Medicine), 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
- Benjamin Guedj is teaching
  - Advanced machine learning (M2, 6h), University College London, United Kingdom
- 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

11.2.2 Supervision

PhD in progress:

- Guillaume Braun, Efficient iterative methods for clustering and matching problems on graphs, [35], January 2020 (defended on December 2022), Christophe Biernacki and Hemant Tyagi
- Eglantine Karle, Dynamic ranking and synchronization on evolving graphs, October 2020, Cristian Preda and Hemant Tyagi
- Axel Potier, sale prediction for low turn-over products, November 2020, Christophe Biernacki, Matthieu Marbac and Vincent Vandewalle
- Luxin Zhang, Agnostic domain adaptation: application to fraud detection, February 2019 (defended on March 2022), Christophe Biernacki, Pascal Germian and Yacine Kessaci
- Filippo Antonazzo, Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach, [34], October 2019 (defended on September 2022), Christophe Biernacki and Christine Keribin
- Wilfried Heyse, Taking into account the temporal structure in the statistical analysis of high-throughput proteomic data, October 2019, Christophe Bauters, Guillemette Marot and Vincent Vandewalle
- Clarisse Boinay, Anomaly detection and change point detection in contextual dynamic asynchronous graphs with applications in OT cybersecurity, December 2021, Christophe Biernacki and Cristian Preda
- Camille Frévent, Contribution to spatial statistics for high-dimensional and survival data, [36] (defended December 2022), Sophie Dabo
- Issam Moindjé, Functional Data Analysis for biomarkers identification on EEG and MEG of feotus and premature, October 2020, Sophie Dabo, Cristian Preda
- François Bassac, Predicting performances from training session data at Decathlon by functional data analysis. October 2022, Cristian Preda

11.2.3 Juries

Christophe Biernacki acted as an examinator for one HdR defense, as reviewer for four PhD defenses (including one at Tel Aviv University in Israël) and as a "follower" of one PhD thesis in Laval University (Canada)

Sophie Dabo-Niang Served as member of jurys of more than 8 PhD, 5 Prize and grants jurys, worldwide.

Cristian Preda Served as reviewer of the HDR of Lionel Cucala (juillet 2022), as a reviewer for the PHD thesis of Cindy Frascolla (Dijon, Juin 2022), as a member of HDR jury for Michael Genin (October 2022), as a examiner for Sfetcu Sorina-Cezarina (Bucharest, September 2022), as a member of the PhD thesis of Camille Frevent (December 2022).

Guillemette Marot served as an examiner for one thesis (Pierre-Emmanuel Desprez, Lille, October 2022).

11.3 Popularization

11.3.1 Interventions

Christophe Biernacki gave an invited talk to the [SystemX seminar] and to the JTR 2022 conference [24]

Sophie Dabo-Niang Girls and Sciences days *Artificial Intelligence Applied to Health*, 4 November 2022, Inria, Lille. Inria Station V on IA applied to industrial problems with Vilogia (Real estate company), October 28th, 2022

12 Scientific production

12.1 Major publications

[1] P. Alquier and B. Guedj. 'Simpler PAC-Bayesian Bounds for Hostile Data'. In: *Machine Learning* (2018). DOI: 10.1007/s10994-017-5690-0. URL: https://hal.inria.fr/hal-01385064.

[2] P. Bathia, S. Iovleff and G. Govaert. 'An R Package and C++ library for Latent block models: Theory, usage and applications'. In: *Journal of Statistical Software* (2016). URL: https://hal.archives-ouvertes.fr/hal-01285610.

[3] C. Biernacki and A. Lourme. 'Unifying Data Units and Models in (Co-)Clustering'. In: *Advances in Data Analysis and Classification* 12.41 (May 2018). URL: 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* 42.5 (2014), pp. 1879–1910. URL: https://hal.archives-ouvertes.fr/hal-00337058.

[5] S. Dabo-Niang, C. Ternynck and A.-F. Yao. 'Nonparametric prediction in the multivariate spatial context'. In: *Journal of Nonparametric Statistics* 28.2 (2016), pp. 428–458. DOI: 10.1080/10485252.2016.01.007. URL: 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 and J. Eeckhoute. 'The logic of transcriptional regulator recruitment architecture at cis-regulatory modules controlling liver functions'. In: *Genome Research* 27.6 (June 2017), pp. 985–996. DOI: 10.1101/gr.217075.116. URL: https://hal.archives-ouvertes.fr/hal-01647846.

[7] G. Letarte, P. Germain, B. Guedj and F. Laviolette. 'Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks'. In: *NeurIPS 2019*. Vancouver, Canada, Dec. 2019. URL: https://hal.inria.fr/hal-02139432.

[8] M. Marbac, C. Biernacki and V. Vandewalle. 'Model-based clustering of Gaussian copulas for mixed data'. In: *Communications in Statistics - Theory and Methods* (Dec. 2016). URL: https://hal.archives-ouvertes.fr/hal-00987760.

[9] C. Preda, Q. Grimonprez and V. Vandewalle. 'Categorical Functional Data Analysis. The cfda R Package'. In: *Mathematics* 9.23 (Dec. 2021), p. 31. DOI: 10.3390/math9233074. URL: https://hal.inria.fr/hal-03515152.

[10] H. Tyagi and J. Vybiral. 'Learning general sparse additive models from point queries in high dimensions'. In: *Constructive Approximation* (Jan. 2019). URL: https://hal.inria.fr/hal-02379404.

12.2 Publications of the year

International journals

[11] A. Chepy, S. Vivier, F. Bray, C. Ternynck, J.-P. Meneboo, M. Figeac, A. Filiot, L. Guilbert, M. Jendoubi, C. Rolando, D. Launay, S. Dubucquoi, G. Marot and V. Sobanski. 'Effects of Immunoglobulins G From Systemic Sclerosis Patients in Normal Dermal Fibroblasts: A Multi-Omics Study'. In: *Frontiers in Immunology*. Frontiers in Immunology 13 (29th June 2022), p. 904631. DOI: 10.3389/fimmu.2022.904631. URL: https://hal.univ-lille.fr/hal-03898588.

[12] M. Cucuringu and H. Tyagi. ‘An extension of the angular synchronization problem to the heterogeneous setting’. In: *Foundations of Data Science* (2022). DOI: 10.3934/fods.2021036. URL: https://hal.inria.fr/hal-03101682.

[13] M. P. B. Gallaugher, C. Biernacki and P. D. Mcnicholas. ‘Parameter-Wise Co-Clustering for High-Dimensional Data’. In: *Computational Statistics* (Oct. 2022). DOI: 10.1007/s00180-022-01289-2. URL: https://hal.science/hal-01862824.

[14] N. Garnier, K. Pollet, M. Fourcot, M. Caplan, G. Marot, J. Goutay, J. Labreuche, F. Soncin, R. Boukherroub, D. Hober, S. Szunerits, J. Poissy and I. Engelmann. ‘[Letter to editor] Altered microRNA expression in severe COVID-19: Potential prognostic and pathophysiological role’. In: *Clinical and Translational Medicine* 12.6 (June 2022). DOI: 10.1002/ctm2.899. URL: https://hal.science/hal-03700328.

[15] Q. Grimonprez, S. Blanck, A. Celisse and G. Marot. ‘MLGL: An R package implementing correlated variable selection by hierarchical clustering and group-Lasso’. In: *Journal of Statistical Software* (2022). URL: https://hal.inria.fr/hal-01857242.

[16] A. Leroy, P. Latouche, B. Guedj and S. Gey. ‘Cluster-Specific Predictions with Multi-Task Gaussian Processes’. In: *Journal of Machine Learning Research* (2022). URL: https://hal.inria.fr/hal-03009276.

[17] A. Leroy, P. Latouche, B. Guedj and S. Gey. ‘MAGMA: Inference and Prediction using Multi-Task Gaussian Processes with Common Mean’. In: *Machine Learning* (6th May 2022). DOI: 10.1007/s10994-022-06172-1. URL: https://hal.inria.fr/hal-02904446.

[18] M. Marbac, M. Sedki, C. Biernacki and V. Vandewalle. ‘Simultaneous semi-parametric estimation of clustering and regression’. In: *Journal of Computational and Graphical Statistics* (2022). URL: https://hal.inria.fr/hal-03090573.

[19] L. Zhang, P. Germain, Y. Kessaci and C. Biernacki. ‘Interpretable Domain Adaptation Using Unsupervised Feature Selection on Pre-trained Source Models’. In: *Neurocomputing* 511 (28th Oct. 2022), pp. 319–336. URL: https://hal.science/hal-03325509.

International peer-reviewed conferences

[20] F. Biggs and B. Guedj. ‘On Margins and Derandomisation in PAC-Bayes’. In: AISTATS 2022 - 25th International Conference on Artificial Intelligence and Statistics. 151. Valencia, Spain, 28th Mar. 2022. URL: https://hal.inria.fr/hal-03282597.

[21] F. Biggs and B. Guedj. ‘Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty’. In: AISTATS. Valencia, Spain, 20th Oct. 2022. URL: https://hal.inria.fr/hal-03953307.

[22] G. Braun, H. Tyagi and C. Biernacki. ‘An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees’. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, United States, 17th July 2022. URL: https://hal.inria.fr/hal-03526257.

[23] B.-E. Chérief-Abdellatif, Y. Shi, A. Doucet and B. Guedj. ‘On PAC-Bayesian reconstruction guarantees for VAEs’. In: 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022. Valencia / Virtual, Spain, 28th Mar. 2022. URL: https://hal.inria.fr/hal-03587178.

National peer-reviewed Conferences

[24] C. Biernacki. ‘Fondamentaux de l’IA et domaines d’application’. In: Journées Techniques Routes 2022. Nantes, France, 10th May 2022. URL: https://hal.inria.fr/hal-03936772.

Conferences without proceedings

[25] C. Biernacki. ‘Impact of Missing Data on Mixtures and Clustering’. In: CMStatistics 2022. London, United Kingdom, 17th Dec. 2022. URL: https://hal.inria.fr/hal-03936786.

[26] C. Biernacki. 'Impact of missing data on mixtures and clustering with illustrations in Biology and Medicine'. In: ASA 2022 Apprentissage Statistiques et Applications. Poitiers, France, 22nd June 2022. URL: https://hal.inria.fr/hal-03936781.

[27] G. Marot. 'Introduction to statistics for omics data'. In: Thematic school FTMS Fourier transform mass spectrometry: treatment of ultra high resolution data. Ambleteuse, France, 14th Nov. 2022. URL: https://hal.inria.fr/hal-03942855.

[28] G. Marot. 'Meta-analysis of RNA-Seq data'. In: Le RNASeq, de la paillasse à l'analyse in silico. Lille, France, 28th June 2022. URL: https://hal.inria.fr/hal-03942820.

[29] G. Marot, Q. Grimonprez, S. Blanck and A. Celisse. 'Variable selection with Multi-Layer Group Lasso'. In: useR! 2022. Virtual, United States, 21st June 2022. URL: https://hal.inria.fr/hal-03942579.

[30] G. Marot, W. Heyse, V. Vandewalle, C. Bauters and F. Pinet. 'Clinical score building from high-throughput proteomic data'. In: ICM/Inria/SCAI workshop "Computational and mathematical approaches for neuroscience". Paris, France, 7th June 2022. URL: https://hal.inria.fr/hal-03942723.

[31] P. Viallard, R. Emonet, P. Germain, A. Habrard, E. Morvant and V. Zantedeschi. 'Intérêt des bornes désintégrées pour la généralisation avec des mesures de complexité'. In: CAp 2022. Vannes, France, 5th July 2022. URL: https://hal.science/hal-03703811.

[32] V. Zantedeschi, P. Viallard, E. Morvant, R. Emonet, A. Habrard, P. Germain and B. Guedj. 'Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound'. In: CAp 2022. Vannes, France, 5th July 2022. URL: https://hal.science/hal-03703804.

[33] L. Zhang, P. Germain, Y. Kessaci and C. Biernacki. 'Interpretable Domain Adaptation for Hidden Subdomain Alignment in the Context of Pre-trained Source Models'. In: 36th AAAI Conférence on Artificial Intelligence. Vancouver, Canada, 22nd Feb. 2022. URL: https://hal.science/hal-03505639.

Doctoral dissertations and habilitation theses

[34] F. Antonazzo. 'Unsupervised learning of huge data sets with limited computed resources'. Université de Lille, 30th Sept. 2022. URL: https://theses.hal.science/tel-03846222.

[35] G. Braun. 'Efficient iterative methods for clustering and matching problems on graphs'. Université de Lille, 6th Dec. 2022. URL: https://hal.science/tel-03889078.

[36] C. Frévent. 'Contribution to spatial statistics for high-dimensional and survival data'. Université de Lille, 2nd Dec. 2022. URL: https://hal.inria.fr/tel-03889127.

Reports & preprints

[37] M. S. Ahmed, M. Attouch, S. Dabo-Niang and M. Ndiaye. *K-nearest neighbors method estimation of regression function for spatial dependent data*. 15th Jan. 2022. URL: https://hal.inria.fr/hal-03527479.

[38] E. Araya, G. Braun and H. Tyagi. *Seeded graph matching for the correlated Wigner model via the projected power method*. 29th Nov. 2022. URL: https://hal.science/hal-03876872.

[39] E. Araya, E. Karlé and H. Tyagi. *Dynamic Ranking and Translation Synchronization*. 29th Nov. 2022. URL: https://hal.science/hal-03876870.

[40] C. Biernacki, J. Jacques and C. Keribin. *A Survey on Model-Based Co-Clustering: High Dimension and Estimation Challenges*. 5th Sept. 2022. URL: https://hal.science/hal-03769727.

[41] G. Braun and H. Tyagi. *Minimax Optimal Clustering of Bipartite Graphs with a Generalized Power Method*. 29th Nov. 2022. URL: https://hal.science/hal-03876871.

[42] E. Clerico, G. Deligiannidis, B. Guedj and A. Doucet. *A PAC-Bayes bound for deterministic classifiers*. 14th Oct. 2022. URL: https://hal.inria.fr/hal-03815146.

[43] S. Dabo-Niang, B. Thiam and T. Verdebout. *Asymptotic efficiency of some nonparametric tests for location on hyperspheres.* 16th Jan. 2022. URL: https://hal.inria.fr/hal-03527763.

[44] C. Frévent, M.-S. Ahmed, S. Dabo-Niang and M. Genin. *Investigating spatial scan statistics for multivariate functional data.* 15th Jan. 2022. URL: https://hal.inria.fr/hal-03527471.

[45] M. Haddouche, B. Guedj and O. Wintenberger. *Optimistic Dynamic Regret Bounds.* 18th Jan. 2023. URL: https://hal.inria.fr/hal-03953310.

[46] S. Iovleff. *A New Probabilistic Representation of the Alternating Zeta Function and a New Selberg-like Integral Evaluation.* 17th Mar. 2022. URL: https://hal.science/hal-03612591.

[47] I.-A. Moindjié, C. Preda and S. Dabo-Niang. *Classification of multivariate functional data on different domains with Partial Least Squares approaches.* 20th Dec. 2022. URL: https://hal.science/hal-03908634.

12.3 Other

Scientific popularization

[48] C. Biernacki. 'Fondamentaux de l'IA et domaines d'application'. In: Journées Techniques Routes. Nantes, France, 10th May 2022. URL: https://hal.inria.fr/hal-03976741.

Educational activities

[49] C. Biernacki. 'Traitement statistique des données manquantes-Part I Introduction to modeling'. Doctoral. France, 10th Mar. 2022. URL: https://hal.science/hal-03505648.

12.4 Cited publications

[50] A. Bahram, M. Deborah, G. Jean-Claude, K. Momoko, F. Hiroyuki and a. et al. 'Single cell classification using statistical learning on mechanical properties measured by mems tweezers.' In: *IEEE 35th International Conference on Micro Electro Mechanical Systems Conference (MEMS 2022)* 0.0 (2022). DOI: 10.1109/MEMS51670.2022.9699466. URL: hhttps://hal.science/hal-03528082/.

[51] S. Sébastien, R. Lisa, T. Ly, V. Eleanor, C. Jean-Luc and a. et al. 'Biomarkers of haemodynamic severity of systemic sclerosis-associated pulmonary arterial hypertension by serum proteome analysis.' In: *Annals of the Rheumatic Diseases* 0.0 (2022). DOI: 10.1136/ard-2022-223237. URL: https://hal.inria.fr/hal-03927525/.