RESEARCH CENTRE

Lille - Nord Europe

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2021 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

Creation of the Project-Team: 2012 January 01

Keywords

Computer sciences and digital sciences

A3.1.4. – Uncertain data

A3.2.3. – Inference

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.5. – Bayesian methods

A3.4.7. – Kernel methods

A5.2. – Data visualization

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, Université de Lille, Professor, HDR]
- Sophie Dabo [Univ Henri Poincaré, Professor]
- Guillemette Marot [Université de Lille, Associate Professor, HDR]
- Vincent Vandewalle [Université de Lille, Associate Professor, HDR]

Post-Doctoral Fellows

- Ernesto Javier Araya Valdivia [Inria]
- Florent Dewez [Inria, until Jan 2021]
- Valentina Zantedeschi [Inria]

PhD Students

- Reuben Adams [University College London]
- Filippo Antonazzo [Inria]
- Felix Biggs [University College London]
- Rajeev Bopche [Inria, until May 2021]
- Guillaume Braun [Institut national de la statistique et des études économiques]
- Theophile Cantelobre [Inria, from Oct 2021]
- Maxime Haddouche [Université de Lille, from Oct 2021]
- Wilfried Heyse [INSERM]
- Eglantine Karlé [Inria]
- Etienne Kronert [Wordline]
- Issam Ali Moindjie [Inria]
- Axel Potier [Groupe Adeo, CIFRE]
- Antonin Schrab [University College London]
- Antoine Vendeville [University College London]
- Luxin Zhang [Wordline, CIFRE]

Technical Staff

- Maxime Brunin [Inria, Engineer, until Mar 2021]
- Ismat Yahia Chaib Draa [Société Alicante à Seclin, Engineer, from Sep 2021]
- Florent Dewez [Inria, Engineer, from Feb 2021 until Apr 2021]

Interns and Apprentices

- Myriam Benbahlouli [Inria, Apprentice, from Oct 2021]
- Claire Devisme [Inria, from Apr 2021 until Aug 2021]
- Ahoua Jean Marc Ehile [Inria, from Mar 2021 until Aug 2021]
- Elias Giraud-Audine [NC, Jul 2021]
- Valentin Kilian [École normale supérieure de Rennes, from May 2021 until Jul 2021]
- Ilyas Lebleu [École Normale Supérieure de Paris, from Apr 2021 until Aug 2021]
- Cecilia Alejandra Rivera Martinez [École Nationale Supérieure d'Arts et Métiers, from Mar 2021 until Aug 2021]
- Seydina Mouhamed Sow [Inria, from Mar 2021 until Aug 2021]

Administrative Assistant

- Anne Rejl [Inria]

External Collaborator

- Alain Celisse [Univ Panthéon Sorbonne, HDR]

2 Overall objectives

2.1 Context

In several respects, modern society has strengthened the need for statistical analysis both from applied and theoretical point 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 respectively as unsupervised learning or supervised learning, even if it is not limited to them or other names exist depending on communities. Somewhere, it pursues the following hope: "more data for better quality and more numerous results".

However, today's data are increasingly complex. They gather mixed type features (for instance continuous data mixed with categorical data), missing or partially missing items (like intervals) and numerous variables (high dimensional situation. As a consequence, the target "better quality and more numerous results" of the previous adage (both words are important: "better quality" and also "more numerous") could not be reached through a somewhat "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.

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 responsability.

6 Highlights of the year

Action Exploratoire (AEx)

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

Acronyme et nom du projet : PATH (PAtient PaThway in the Hospital environment)

Collaboration : CHU Lille and Faculty of Medicine (METRICS team)
Researchers involved : Jean-Baptiste Beuscart, Grégoire Ficheur, Emmanuel Chazard, Michaël, Genin, Antoine Lamer, Génia Babykina, Cyrielle Dumont.

European healthcare systems are faced with multiple challenges, including an aging population, an increase in chronic diseases and patients with multiple illnesses, and limited financial and human resources. The response to these challenges relies in particular on the organization of care into care pathways, justified by the scientific literature and supported in France by political orientations. The analysis of care pathways and their adequacy to needs and resources has thus become a major scientific and administrative challenge. Although the numerical data available for this purpose are rapidly increasing, the methods and statistical tools available to researchers and health authorities remain limited and inefficient. PATH proposes to develop statistical methods for the construction/analysis of the patient

pathway through two applications dealing with the re-hospitalization of the elderly and post-operative complications.

Inria London. Benjamin Guedj has led the emerging Inria London Programme since 2019. The partnership involves Inria and University College London (UCL) as of February 1st, 2021 and the official kickoff.

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

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 MASSICCC

Name: Massive Clustering with Cloud Computing

Keywords: Statistic analysis, Big data, Machine learning, Web Application

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

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

URL: https://massiccc.lille.inria.fr

Contact: Christophe Biernacki

7.1.3 cfda

Name: Categorical functional data analysis

Keyword: Functional data

Functional Description: The R package cfda performs: - descriptive statistics for categorical functional data - dimension reduction end 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

7.1.4 PyRotor

Name: Python Route Trajectory Optimiser

Keywords: Optimization, Machine learning, Trajectory Modeling

Scientific Description: PyRotor is a Python implementation of the trajectory optimisation method introduced in the paper: "An end-to-end data-driven optimisation framework for constrained trajectories"

The method proposes trajectories optimizing a given criterion. Unlike classical approaches (such as optimal control), the method is based on the information contained in the available data. This permits to restrict the search area to a neighborhood of the observed trajectories and incorporates the correlations estimated from the data. This is achieved by means of a regularization term in the cost function. An iterative approach is also developed to verify additional constraints.

Functional Description: PyRotor leverages available trajectory data to focus the search space and to estimate some properties which are then incorporated in the optimisation problem. This constraints in a natural and simple way the optimisation problem whose solution inherits realistic patterns from the data. In particular PyRotor does not require any knowledge on the dynamics of the system.

News of the Year: Methodology development and implementation of the first results

URL: https://pypi.org/project/pyrotor/

Publication: hal-03024720

Contact: Florent Dewez

Participants: Florent Dewez, Benjamin Guedj, Arthur Talpaert, Vincent Vandewalle

7.1.5 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 described in Marbac et al. (2020) <arXiv:2012.14159> are implemented dans ce package R

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

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

Contact: Matthieu Marbac-Lourdelle

7.1.6 MPAGenomics

Name: Multi-Patient Analysis of Genomic markers

Keywords: Segmentation, Genomics, Marker selection, Biostatistics

Scientific Description: MPAgenomics is an R package for multi-patients analysis of genomics markers. It enables to study several copy number and SNP data profiles at the same time. It offers wrappers from commonly used packages to offer a pipeline for beginners in R. It also proposes a special way of choosing some crucial parameters to change some default values which were not adapted in the original packages. For multi-patients analysis, it wraps some penalized regression methods implemented in HDPenReg.

Functional Description: MPAgenomics provides functions to preprocess and analyze genomic data. It is devoted to: (i) efficient segmentation and (ii) genomic marker selection from multi-patient copy number and SNP data profiles.

Release Contributions: The initial version of MPAGenomics was relying on CGHSeg R package, which was providing fast segmentation. However, CGHSeg was not maintained. As it was not developped by MODAL team, we did not want to maintain it. In order to let MPAGenomics on the CRAN, Samuel Blanck created a new version of MPAGenomics without dependence to CGHSeg (version 1.2.3) and offered a complete version with dependence to CGHseg (which is of quality, even if not maintained) at https://github.com/sblanck/MPAgenomics.

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

Contact: Samuel Blanck

Participants: Guillemette Marot, Quentin Grimonprez, Samuel Blanck

Partner: Université de Lille

7.1.7 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

7.1.8 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

7.1.9 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

Contact: Sophie Dabo

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: Model-based Co-clustering for Ordinal Data of Different Dimensions

Participants: Christophe Biernacki.

This work has been motivated by a psychological survey on women affected by a breast tumor. Patients replied at different moments of their treatment to questionnaires with answers on ordinal scale. The questions relate to aspects of their life called dimensions. To assist the psychologists in analyzing the results, it is useful to emphasize a structure in the dataset. The clustering method achieves that by creating groups of individuals that are depicted by a representative of the group. From a psychological position, it is also useful to observe how questions may be grouped. This is why a clustering should also be performed on the features, which is called a co-clustering problem. However, gathering questions that are not related to the same dimension does not make sense from a psychologist stance. Therefore, the present work corresponds to perform a constrained co-clustering method aiming to prevent questions from different dimensions from getting assembled in a same column-cluster. In addition, evolution of co-clusters along time has been investigated. The method relies on a constrained Latent Block Model embedding a probability distribution for ordinal data. Parameter estimation relies on a Stochastic EM-algorithm associated to a Gibbs sampler, and the ICL-BIC criterion is used for selecting the numbers of co-clusters. The resulting work has been accepted in an international journal in 2019 and the related R package ordinalClust has been accepted this year in another international journal [26].

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

8.2 Axis 1: Gaussian-based Visualization of Gaussian and non-Gaussian Model-based Clustering

Participants: Christophe Biernacki, Vincent Vandewalle.

A generic method is introduced to visualize in a Gaussian-like way, and onto R^2, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have a thorough overview of the potentially complex clustering result. An entropic measure allows us to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis. This work has been published last year in an international journal [85], and has been presented in a conference [40]. This is a joint work with Matthieu Marbac from ENSAI.

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 and a talk to a conference has been given [38]. A more general talk on missing data and its impact on mixtures and clustering has also been given this year in a workshop [39].

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 [24] and has been presented in an national conference [34].

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 [54], a short version has been accepted in a book of short papers associated to an international conference [47] and has to a talk in an international workshop and also to a seminar .

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

8.6 Axis 1: Regularized spectral methods for clustering signed networks

Participants: Hemant Tyagi.

We study the problem of k-way clustering in signed graphs. Considerable attention in recent years has been devoted to analyzing and modeling signed graphs, where the affinity measure between nodes takes either positive or negative values. Recently, Cucuringu et al. [CDGT 2019] proposed a spectral method, namely SPONGE (Signed Positive over Negative Generalized Eigenproblem), which casts the clustering task as a generalized eigenvalue problem optimizing a suitably defined objective function. This approach is motivated by social balance theory, where the clustering task aims to decompose a given network into disjoint groups, such that individuals within the same group are connected by as many positive edges as possible, while individuals from different groups are mainly connected by negative edges. Through extensive numerical simulations, SPONGE was shown to achieve state-of-the-art empirical performance. On the theoretical front, [CDGT 2019] analyzed SPONGE and the popular Signed Laplacian method under the setting of a Signed Stochastic Block Model (SSBM), for $k = 2$ equal-sized clusters, in the regime where the graph is moderately dense. In this work, we build on the results in [CDGT 2019] on two fronts for the normalized versions of SPONGE and the Signed Laplacian. Firstly, for both algorithms, we extend the theoretical analysis in [CDGT 2019] to the general setting of $k \geq 2$ unequal-sized clusters in the moderately dense regime. Secondly, we introduce regularized versions of both methods to handle sparse graphs – a regime where standard spectral methods underperform – and provide theoretical guarantees under the same SSBM model. To the best of our knowledge, regularized spectral methods have so far not been considered in the setting of clustering signed graphs. We complement our theoretical results with an extensive set of numerical experiments on synthetic data.

This is joint work with Mihai Cucuringu (University of Oxford, United Kingdom), Apoorv Vikram Singh (NYU), Deborah Sulem (University of Oxford, United Kingdom). It was initiated when Apoorv Vikram Singh visited the MODAL team to work with Hemant Tyagi from Oct 2019-Jan 2020. This has now been accepted for publication in the journal: *Journal of Machine Learning Research* [12]. A summary of the results was presented at the GCLR (Graphs and more Complex structures for Learning and Reasoning) workshop at AAAI 2021.

8.7 Axis 1: Dynamic Ranking with the BTL Model: A Nearest Neighbor based Rank Centrality Method

Participants: Eglantine Karlé, Hemant Tyagi.

Many applications such as recommendation systems or sports tournaments involve pairwise comparisons within a collection of n items, the goal being to aggregate the binary outcomes of the comparisons in order to recover the latent strength and/or global ranking of the items. In recent years, this problem has received significant interest from a theoretical perspective with a number of methods being proposed, along with associated statistical guarantees under the assumption of a suitable generative model. While these results typically collect the pairwise comparisons as one comparison graph G, however in many applications-such as the outcomes of soccer matches during a tournamentthe nature of pairwise outcomes can evolve with time. Theoretical results for such a dynamic setting are relatively limited compared to the aforementioned static setting. We study in this paper an extension of the classic BTL (Bradley-Terry-Luce) model for the static setting to our dynamic setup under the assumption that the probabilities of the pairwise outcomes evolve smoothly over the time domain [0, 1]. Given a sequence of comparison graphs $(G'_t)_{t' \in T}$ on a regular grid $T \subset [0,1]$, we aim at recovering the latent strengths of the items $w^*_t \in R^n$ at any time $t \in [0,1]$. To this end, we adapt the Rank Centrality method-a popular spectral approach for ranking in the static case-by locally averaging the available data on a suitable neighborhood of t. When $(G'_t)_{t' \in T}$ is a sequence of Erdös-Renyi graphs, we provide non-asymptotic ℓ_2 and ℓ_∞ error bounds for estimating w^*_t which in particular establishes the consistency of this method in terms of n,

and the grid size |T |. We also complement our theoretical analysis with experiments on real and synthetic data. This PhD work [65] has been submitted to a journal and is currently under review.

8.8 Axis 1: An extension of the angular synchronization problem to the heterogeneous setting

Participants: Hemant Tyagi.

Given an undirected measurement graph $G = ([n], E)$, the classical angular synchronization problem consists of recovering unknown angles $\theta_1, \ldots, \theta_n$ from a collection of noisy pairwise measurements of the form $(\theta_i - \theta_j) \bmod 2\pi$, for each $\{i, j\} \in E$. This problem arises in a variety of applications, including computer vision, time synchronization of distributed networks, and ranking from preference relationships. In this paper, we consider a generalization to the setting where there exist k unknown groups of angles $\theta_{l,1}, \ldots, \theta_{l,n}$, for $l = 1, \ldots, k$. For each $\{i, j\} \in E$, we are given noisy pairwise measurements of the form $\theta_{\ell,i} - \theta_{\ell,j}$ for an *unknown* $\ell \in \{1, 2, \ldots, k\}$. This can be thought of as a natural extension of the angular synchronization problem to the heterogeneous setting of multiple groups of angles, where the measurement graph has an unknown edge-disjoint decomposition $G = G_1 \cup G_2 \ldots \cup G_k$, where the G_i's denote the subgraphs of edges corresponding to each group. We propose a probabilistic generative model for this problem, along with a spectral algorithm for which we provide a detailed theoretical analysis in terms of robustness against both sampling sparsity and noise. The theoretical findings are complemented by a comprehensive set of numerical experiments, showcasing the efficacy of our algorithm under various parameter regimes. Finally, we consider an application of bi-synchronization to the graph realization problem, and provide along the way an iterative graph disentangling procedure that uncovers the subgraphs G_i, $i = 1, \ldots, k$ which is of independent interest, as it is shown to improve the final recovery accuracy across all the experiments considered.

This is joint work with Mihai Cucuringu (University of Oxford, United Kingdom) and has been accepted for publication in the journal *Foundations of Data Science* [13].

8.9 Axis 1&2: Clustering on Multilayer Graphs with Missing Values

Participants: Christophe Biernacki, Guillaume Braun, Hemant Tyagi.

Multilayer graphs clustering have gained increasing interest this last decade due to numerous applications in various fields. Several clustering methods have been proposed, but they rely all on the assumption that the network is fully observed. We propose a statistical framework to handle nodes that are missing on some layers as well as a method to estimate the model parameters and to impute missing edge values.

This work has been published in an international conference [41].

8.10 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 hyper-parameter 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 been submited to an international conference and is currently available as a preprint [57].

8.11 Axis 2: Denoising modulo samples: k-NN regression and tightness of SDP relaxation

Participants: Hemant Tyagi.

Many modern applications involve the acquisition of noisy modulo samples of a function f, with the goal being to recover estimates of the original samples of f. For a Lipschitz function $f : [0,1]^d \to \mathbb{R}$, suppose we are given the samples $y_i = (f(x_i) + \eta_i) \bmod 1; \quad i = 1,\ldots,n$ where η_i denotes noise. Assuming η_i are zero-mean i.i.d Gaussian's, and x_i's form a uniform grid, we derive a two-stage algorithm that recovers estimates of the samples $f(x_i)$ with a uniform error rate $O((\frac{\log n}{n})^{\frac{1}{d+2}})$ holding with high probability. The first stage involves embedding the points on the unit complex circle, and obtaining denoised estimates of $f(x_i) \bmod 1$ via a kNN (nearest neighbor) estimator.

Recently, Cucuringu and Tyagi proposed an alternative way of denoising modulo 1 data which works with their representation on the unit complex circle. They formulated a smoothness regularized least squares problem on the product manifold of unit circles, where the smoothness is measured with respect to the Laplacian of a proximity graph G involving the x_i's. This is a nonconvex quadratically constrained quadratic program (QCQP) hence they proposed solving its semidefinite program (SDP) based relaxation. We derive sufficient conditions under which the SDP is a tight relaxation of the QCQP. Hence under these conditions, the global solution of QCQP can be obtained in polynomial time.

This is joint work with Michael Fanuel (KU Leuven). It is has been accepted for publication in the journal *Information and Inference: A journal of the IMA* [17].

8.12 Axis 2: Error analysis for denoising smooth modulo signals on a graph

Participants: Hemant Tyagi.

In many applications, we are given access to noisy *modulo* samples of a smooth function with the goal being to robustly unwrap the samples, i.e. to estimate the original samples of the function. In a recent work, Cucuringu and Tyagi proposed denoising the modulo samples by first representing them on the unit complex circle and then solving a smoothness regularized least squares problem – the smoothness measured w.r.t. the Laplacian of a suitable proximity graph G – on the product manifold of unit circles. This problem is a quadratically constrained quadratic program (QCQP) which is nonconvex, hence they proposed solving its *sphere-relaxation* leading to a trust region subproblem (TRS). In terms of theoretical guarantees, ℓ_2 error bounds were derived for (TRS). These bounds are however weak in general and do not really demonstrate the denoising performed by (TRS).

In this work, we analyse the (TRS) as well as an unconstrained relaxation of (QCQP). For both these estimators we provide a refined analysis in the setting of Gaussian noise and derive noise regimes where they provably denoise the modulo observations w.r.t. the ℓ_2 norm. The analysis is performed in a general setting where G is any connected graph.

This work has been accepted for publication in the journal *Applied and Computational Harmonic Analysis* [27].

8.13 Axis 2: Recovering Hölder smooth functions from noisy modulo samples

Participants: Hemant Tyagi.

In signal processing, several applications involve the recovery of a function given noisy modulo samples. The setting considered in this paper is that the samples corrupted by an additive Gaussian noise are wrapped due to the modulo operation. Typical examples of this problem arise in phase unwrapping problems or in the context of self-reset analog to digital converters. We consider a fixed design setting where the modulo samples are given on a regular grid. Then, a three stage recovery strategy is proposed to recover the ground truth signal up to a global integer shift. The first stage denoises the modulo samples by using local polynomial estimators. In the second stage, an unwrapping algorithm is applied to the denoised modulo samples on the grid. Finally, a spline based quasi-interpolant operator is used to yield an estimate of the ground truth function up to a global integer shift. For a function in Hölder class, uniform error rates are given for recovery performance with high probability. This extends recent results obtained by Fanuel and Tyagi for Lipschitz smooth functions wherein kNN regression was used in the denoising step.

This is joint work with Michaël Fanuel (CRIStAL, Universite de Lille) and was presented in an invited session on Computational Sampling in an international conference (ASILOMAR 2021). The paper [63] will be appearing in the proceedings of the conference.

8.14 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, Painleve), Thomas Verdebout (ULB, Belgium). This work has been been submitted for publication [62].

8.15 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 (Uviversity Sidi Bel Abbes, Algeria), Mamadou Ndiaye (UCAD, Senegal). This work is under revision. [53].

8.16 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). This work has been published in the book Geostatistical Functional Data Analysis: Theory and Methods. J. Mateu and R. Giraldo (Eds). John Wiley and Sons, Chichester, UK. ISBN: 978-1-119-38784-8

8.17 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). This work has been published in the book Geostatistical Functional Data Analysis: Theory and Methods. J. Mateu and R. Giraldo (Eds). John Wiley and Sons, Chichester, UK. ISBN: 978-1-119-38784-8

8.18 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.

This work has been published in the book Geostatistical Functional Data Analysis: Theory and Methods. J. Mateu and R. Giraldo (Eds). John Wiley and Sons, Chichester, UK. ISBN: 978-1-119-38784-8 [49].

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

Participants: Sophie Dabo-Niang.

This paper introduces the R package HDSpatialSca 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). A R package name HDSpatialSca has been developped, a related article is under revision in R journal. [64].

8.20 Axis 3: Categorical functional data analysis

Participants: Cristian Preda, Quentin Grimonprez, Vincent Vandewalle.

We have developed the methodology to visualize, perform dimension reduction and extract features from categorical functional data. For this, the *cfda* R package has been developed and added to CRAN repository. A paper presenting the features of the cfda R package (7.1.3) with an application to care data (clustering of patient paths) is published in [9]. The cfda R package has been presented to the ERCIM [77] and SFDS [78].

8.21 Axis 3: Clustering categorical functional data

Participants: Cristian Preda, Vincent Vandewalle.

The objective of this research direction was: (i) to propose possible modelling approaches of categorical functional data and (ii) to investigate the identifiability problem of such models. A first modelling framework is to consider that an observed functional data path represents a sample path of Markov process and thus n sample paths come from several, say K, different processes. Consequently, we have here a mixture of K different Markov processes. A second modelling framework is to consider that the observed sample path come from several semi-Markov processes. The parameter estimation is obtained through techniques based on the EM algorithm, while the selection of the number of classes is based on information criteria. An important problem is to determine the class membership for each sample paths, but our main concern is related to the identifiability problem. The identifiability of this type of models cannot be obtained in general, but only by imposing restrictions on the parameters of the model, cf. [87, 88]. Our work in progress is related to find sufficiently general conditions that guarantee this identifiability.

8.22 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. One of the main challenge is to take into account the variations of the biomarkers according to the time (several measurement times), in order to improve the understanding of biological mechanisms involved on LVR or survival of the patient.

By selecting 46 proteins significantly associated to long-term survival in Cox models we have identified 2 groups of patients (one group with high risk and the other with lower risk). Network analysis identified common pathways from the 46 proteins related to cell death and survival and cell-to-cell communication. This work as been submitted for possible publication in an international journal, and has been presented in a national conference [42].

We are now investigating the possibilities to take into account the temporal structure and the high dimension of the data. The aim of this work is to jointly model the temporal structure of all the proteins and the long-term survival of the patients. The main challenges of this work are both the number of proteins and the repeated measurements and the strategy to address these challenges are to introduce the information of known groups of proteins defined by the GO categories (known categories of proteins which are part of a same biological function). This work could lead to the identification of new biomarkers for heart failure.

This is a joint work with Florence Pinet from INSERM.

8.23 Axis 4: Contribution to the nutritional transition

Participants: Wilfried Heyse.

The nutritional transition of a country is characterized by a shift from a traditional, generally plant-rich diet to a meat-rich diet. In the last half century, several countries have experienced significant economic development and have seen their food supply increase. The aim of this work was to identify similarities in food transitions across countries of the world over the past 60 years. The food availability data of 171 countries (public FAOSTATS) gather the informations total food availability, as well as per capita animal and plant products per year, for each country over the period 1961-2018. In order to identify transition patterns, we used unsupervised clustering (hierarchical clustering) analyses which led us to identify 5 distinct clusters with different food transitions patterns. In conclusion, between 1961 and 2018, total food availability increased overall, but with regional disparities. Several transition patterns were identified, characterized by a fairly marked increase in the availability of plant products and to a lesser degree meat. The level of economic development and geographical location are strong indicators of food transition patterns in the world. This work has been presented in a national peer review conference [35].

8.24 Axis 4: Identification of a new master regulator through transcriptomics and epigenomics data analysis

Participants: Guillemette Marot.

Thanks to a suitable analysis of RNA-seq and ChIP-seq data, our collaborators have identified a master regulator which controls asexual cell cycle division patterns in Toxoplasma gondii. Guillemette Marot essentially participated to the statistical analysis of the transcriptomic and epigenomic data of the project, bringing her expertise on empirical bayesian approaches useful to obtain and interpret the results. This joint work with Mathieu Gissot (PI), H. Touzet and P. Pericard for the bioinformatics part, and other collaborators for the biological part, was published in Nature Communications [23].

8.25 Axis 4: Reject Inference Methods in Credit Scoring

Participants: Christophe Biernacki, Adrien Ehrhardt, Philippe Heinrich, Vincent Vandewalle.

The granting process of all credit institutions rejects applicants having a low credit score. Developing a scorecard, i.e. a correspondence table between a client's characteristics and his score, requires a learning dataset in which the target variable good/bad borrower is known. Rejected applicants are de facto excluded from the process. This biased learning population might have deep consequences on the scorecard relevance. Some works, mostly empirical ones, try to exploit rejected applicants in the scorecard building process. This work proposes a rational criterion to evaluate the quality of a scoring model for the existing Reject Inference methods and dig out their implicit mathematical hypotheses. It is shown that, up to now, no such Reject Inference method can guarantee a better credit scorecard. These conclusions are illustrated on simulated and real data from the french branch of Crédit Agricole Consumer Finance (CACF). This work is now published in an international journal [16].

This is a joint work with Sébastien Beben of Crédit Agricole Consumer Finance.

8.26 Axis 4: Usability study

Participants: Vincent Vandewalle.

Since 2018, Vincent Vandewalle is working with Alexandre Caron and Benoît Dervaux (ULR 2694 – METRICS) on issues of estimating the number of problems and the value of information in the field of usability. Based on usability study of a medical device the objective is to determine the number of possible problems linked to the use of a medical device (e.g. insulin pump) as well as their respective occurrence probabilities. Estimating this number and the different probabilities is essential to determine whether or not an additional usability study should be conducted, and to determine the number of users to be included in this study to maximize the expected benefits.

The discovery process can be modeled by a binary matrix, a matrix whose number of columns depends on the number of defects discovered by users. In this framework, they have proposed a probabilistic modeling of this matrix. They have included this modeling in a Bayesian framework where the number of problems and the probabilities of discovery are considered as random variables. It shows the interest of the approach compared to the approaches proposed in the state of the art in usability. The approach beyond point estimation also makes it possible to obtain the distribution of the number of problems and their respective probabilities given the discovery matrix.

The proposed model published in last year [89] also allows to implement an approach aiming at measuring the value of additional information in relation to the discovery process. In this framework, they have written a second paper accepted for publication in Value in Health. They are also developing the R package useval available soon.

8.27 Axis 4: Artificial intelligence for aviation

Participants: Florent Dewez, Benjamin Guedj, Arthur Talpaert, Vincent Vandewalle.

Since November 2018, Benjamin Guedj and Vincent Vandewalle have been participating in the European PERF-AI project (European PERF-AI project: Enhance Aircraft Performance and Optimization through the utilization of Artificial Intelligence) in partnership with the company Safety Line. In particular, using data collected during flights involves developing Machine Learning models to optimize the aircraft's trajectory concerning fuel consumption, for example. In this context they have hired Florent Dewez (post-doctoral researcher) and Arthur Talpaert (engeneer).

The article [86] has been published last year. It explains how, using flight recording data, it is possible to implement learning models on variables that have not been directly observed, and in particular to predict the drag and lift coefficients as a function of the angle and speed of the aircraft.

A second article is in revision about the optimization of the aircraft's trajectory based on a consumption model learned from the data.

The originality of the approach consists in decomposing the trajectory on a functional basis, and thus carrying out the optimization on the coefficients of the decomposition on this basis, rather than approaching the problem from the angle of optimal control. Furthermore, to guarantee compliance with aeronautical constraints, we have proposed an approach penalized by a deviation term from reference flights. A generic Python module (PyRotor) to solve such optimization problems in conjunction with the proposed approach has been developed.

8.28 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.

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

8.29 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 submited to an international journal [75].

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

8.30 Other: Projection Under Pairwise Control

Participants: Christophe Biernacki.

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

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

8.31 Other: On the Local and Global Properties of the Gravitational Spheres of Influence

Participants: Christophe Biernacki.

We revisit the concept of sphere of gravitational activity, to which we give both a geometrical and physical meaning. This study aims to refine this concept in a much broader context that could, for instance, be applied to exo-planetary problems (in a Galactic stellar disc-StarPlanets system) to define a first order "border" of a planetary system. The methods used in this paper rely on classical Celestial Mechanics and develop the equations of motion in the framework of the 3-body problem (e.g. Star-Planet-Satellite System. We start with the basic definition of planet's sphere of activity as the region of space in which it is feasible to assume a planet as the central body and the Sun as the perturbing body when computing perturbations of the satellite's motion. We then investigate the geometrical properties and physical meaning of the ratios of Solar accelerations (central and perturbing) and planetary accelerations (central and perturbing), and the boundaries they define. We clearly distinguish throughout the paper between the sphere of activity, the Chebotarev sphere (a particular case of the sphere of activity), Laplace sphere, and the Hill sphere. The last two are often wrongfully thought to be one and the same. Furthermore, taking a closer look and comparing the ratio of the star's accelerations (central/perturbing) to that of the planetary acceleration (central/perturbing) as a function of the planeto-centric distance, we have identified different dynamical regimes which are presented in the semi-analytical analysis.

This a joint work with Damya Souami from Observatoire de Paris and with Jacky Cresson from Université de Pau et des Pays de l'Adour.

8.32 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 ton an international conference "The 35th International Conference on Micro Electro Mechanical Systems", on January 2022. [58].

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.

8.33 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).

8.34 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).

8.35 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.

Accepted at the Bayesian Deep Learning workshop at NeurIPS 2021.

8.36 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.

8.37 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.

8.38 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.

Accepted at AISTATS 2022.

8.39 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.

8.40 Axis 4: Covid-19 and AI: Unexpected Challenges and Lessons

Participants: Benjamin Guedj.

On May 21st, 2021, we held the webinar "Covid-19 and AI: unexpected challenges and lessons". This short note presents its highlights.

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

Participants: Benjamin Guedj, Antoine Vendeville.

In this paper 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.

8.42 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.

8.43 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.

8.44 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.

8.45 Axis 1: Online k-means Clustering

Participants: Benjamin Guedj.

Abstract on AISTATS 2021 proceedings.

8.46 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 paper proposes a novel algorithm to automatically and sequentially learn principal curves from data streams. We show that our procedure is supported by regret bounds with optimal sublinear remainder terms. A greedy local search implementation (called slpc, for sequential learning principal curves) that incorporates both sleeping experts and multi-armed bandit ingredients is presented, along with its regret computation and performance on synthetic and real-life data.

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.

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

Ordoclic company

Participants: Christophe Biernacki, Cristian Preda.

Duration : April 1st - August 31 2021.

ORDOCLIC is a private company the main activity is the electronic medical prescription. It wishes to take advantage of the richness of its offer to build the first French platform of intelligent prescription taking into account the sanitary state in real time on the territory: to suggest treatments according to the season, the clusters but also to alert on epidemics in connection with the anonymized contents of prescriptions. In this projet, MODAL helped for contextualised prescription aid for infectious diseases in general practice (influenza, gastroenteritis, SARS, chickenpox, etc.). Claire Devisme (Polytech'Lille student) worked for 5 months as an internship on the subject.

COLAS company

Participants: Christophe Biernacki.

COLAS is a world leader in the construction and maintenance of transport infrastructure. This bilateral contract aims at classifying mixed data obtained with sensors coming from a study of the aging of road surfacing. The challenge is to deal with many missing (sensors failures) and correlated data (sensors proximity). This 2nd contract with COLAS finished in 2021.

PAY-BACK company

Participants: Christophe Biernacki.

PAY-BACK Group is an audit firm specializing in the analysis and reliability of transactions. This bilateral contract aims at predicting store sales both from past sales (times series) and also by exploiting external covariates (of different types). The proposed solution is based on the MixtComp software. In 2021, PAY-BACK implemented the MixtComp software in its own information system.

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.

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).

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.

10 Partnerships and cooperations

10.1 European initiatives

10.1.1 FP7 & H2020 projects

H2020 FAIR

Participants: Guillemette Marot, Maxime Brunin.

- Acronym: FAIR
- Project title: Flagellin aerosol therapy as an immunomodulatory adjunct to the antibiotic treatment of drug-resistant bacterial pneunomia
- Coordinator: JC Sirard (Inserm, CIIL)
- 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 bililleplatform, and supervises and participates to data analysis of omic data. She also contributes as a researcher, following Maxime Brunin's preliminary work, for the development of a tool to facilitate multi-omics data integration.

10.2 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.

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.

COVIDOM project

Participants: Christophe Biernacki, Vincent Vandewalle.

- **Abstract:** During the 1st lockdown in France, Christophe Biernacki supervised a task force composed of three Inria research teams (MODAL, STATIFY, TAU) for analysing data coming from the medical database COVIDOM of AP-HP concerning suspected COVID-19 patients. This project was included in the overall national Inria "mission COVID" initiative.

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.

10.2.1 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)

TheraSCUD2022

Participants: Guillemette Marot, Maxime Brunin.

- **Acronym:** TheraSCUD2022
- **Project title:** Targeting the IL-20/IL-22 balance to restore pulmonary, intestinal and metabolic homeostasis after cigarette smoking and unhealthy diet
- **Coordinator:** P. Gosset (Institut Pasteur de Lille)
- **Duration:** 42 months (2017–2021)
- **Partners:** CIIL Institut Pasteur de Lille and UMR 1019 INRA Clermont-Ferrand
- **Abstract:** The TheraSCUD2022 project studies inflammatory disorders associated with cigarette smoking and unhealthy diet (SCUD). Guillemette Marot is involved in this ANR project as head of bilille platform, and has supervised Maxime Brunin on integration of omic data. More information on the ANR website

10.2.2 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).

10.2.3 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

10.3 Regional initiatives

10.3.1 bilille, the bioinformatics platform of Lille

Participants: Guillemette Marot, Maxime Brunin.

The bioinformatics platform of Lille, bilille, is part of UMS 2014/US 41 PLBS (Plateformes Lilloises en Biologie Santé). Guillemette Marot is the scientific head of the platform. In 2021, Inria employed during 3 months Maxime Brunin as engineer for this platform: he participated to the development of visCorVar, a tool which facilitates multi-block analysis for statistical integration of omics data. This was a collaboration needed for the transition between the ANR TheraSCUD2022 project and the european H2020 project FAIR.

Collaborations of the year linked to bilille

Participants: Guillemette Marot.

Guillemette Marot has supervised the data analysis part or support in biostatistics tools testing 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):

- Virology lab ULR3610, I. Engelmann
- Life Imaging Platform PLBS UMS2014 – US41, R Viard
- Canther UMR9020 - U1277, E. Bonnelye

10.3.2 ONCOLILLE

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

Institute for Interdisciplinary Research in Cancerology, named textcolorblueONCOLille, created on January 1, 2020. The objective of the institute is to develop interdisciplinary research associating biology, physics, chemistry, mathematics, bioinformatics, economics, health technologies, human and social sciences by developing fundamental research and strong translational/pre-clinical research (development of alternative and original study models) in order to move towards transfer to the clinic (clinical trials, new molecules). ONCOLille is supported by the University of Lille, Inserm, CNRS, the Lille University Hospital, the Oscar Lambret Center CRCC, the Pasteur Institute of Lille (IPL), and the Lille Cancer Research Institute (IRCL), as well as strong support from the State, the Hauts-de-France region, the MEL (Lille European Metropolis) and the ERDF (European Regional Development Fund). ONCOLille researchers come from 7 laboratories (CANTHER, PHYCELL, ONCOTHAI, LIMMS, SCALab, LPP, LEM) covering all the disciplines of the institute (note 2 laboratories associated with ONCOLille; PRISM and UGSF). The central research theme for all the ONCOLille teams is resistance to treatment and tumor dormancy. This resistance, which can take many forms, is very often the cause of the failure of current therapies. It is this resistance, through knowledge of the mechanisms, that researchers and clinicians will seek to combat and/or divert in order to make the tumor sensitive to treatment again, to propose new treatments through the development of new therapeutic molecules and/or the proposal of new drug combinations and the development of new technologies to treat cancers.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Hemant Tyagi is the organizer of the MODAL team scientific seminar.
- Christophe Biernacki has been president of the scientific comitee of JdS 2021, the annual national meeting the French staticial society (SFdS).
- Sophie Dabo is co-chair of the group Statistics, applied math and computer science of Pan-African Scientific Research Council, funded by Princeton University (USA)

11.1.2 Scientific events: selection

Member of the conference program committees

- Christophe Biernacki has been member of the program committee of the CLADAG 2021 conference.
- Cristian Preda has been a member of the scientific program committee of the Young Researchers Workshop of the SPSR, (Romanian Society of Statistics and Probabilitiy)
- Sophie Dabo has ben member of the scientific program committee of the 14th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2021)
- Benjamin Guedj has been member of programme committees (reviewer) for ALT, AISTATS, ICML, ICLR, NeurIPS, COLT.
- Benjamin Guedj has served as Area Chair for NeurIPS 2021.

11.1.3 Journal

Member of the editorial boards

- Cristian Preda is an Associate Editor for *Methodology and Computing in Applied Probability* journal, Springer.
- Christophe Biernacki is an Associate Editor of the North-Western European Journal of Mathematics (NWEJM).
- Sophie Dabo is an associate editor of Revista Colombiana de Estadística Journal Of Statistical Modeling and Analytics, Journal of Nonparametric Statistics
- Benjamin Guedj is a member of the Editorial Boards of the journals Information and Inference, and Data-Centric Engineering.

Reviewer - reviewing activities

- Hemant Tyagi has reviewed for the following journals during 2021: IEEE Open Journal of Signal Processing, IEEE Transactions on Signal Processing, Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Information and Inference.
- Hemant Tyagi has reviewed for the following conferences in 2021: ICML, ICLR, NeurIPS.
- Christophe Biernacki has reviewed for the Cap2021 (Conférence sur l'Apprentissage Automatique) and also for several journals (LSTA, ADAC,JCGS, CSDA, CLAS, AIRE).
- Sophie Dabo has reviewed the following journals : JRSS B, JASA, JNP, Spatial Statistics
- Cristian Preda has reviewed for the following journals durin 2021 : Methodology and Computing in Applied Probability, Electronic Jurnal of Statistics, Australian and New Zeeland Journal of Statistics.
- Sophie Dabo has reviewed the following journals : JRSS B, JASA, JNP, Spatial Statistics

11.1.4 Invited talks

- Hemant Tyagi gave an invited talk in the session "Computational Sampling" at the ASILOMAR Conference on Signals, Systems and Computers, 2021.
- Christophe Biernacki has been invited to several conferences and workshop for giving a talk: MHC2021 [39], MIMO 2021 [40], a conference on modeling and forecasting in Poland [38].
- Cristian Preda gave an invited talk to Romanian Academy Research Conference, CCSAR 2021.

- Sophie Dabo gave an invited talk to "STOCHASTIC GEOMETRY DAYS, 9th Edition, November 15th-19th, 2021, Dunkerque"
- Sophie Dabo gave an invited talk to monthly INRIA LIRIMA seminar, 2021.
- Benjamin Guedj gave a number of invited talks, mostly in the UK.

11.1.5 Scientific expertise

- Christophe Biernacki has acted as an expert for HCERES for the mathematics laboratory of the Brest-Vannes University. He acted also as an expert for helping the Reims University to prepare its future mathematics laboratory HCERES evaluation.
- Benjamin Guedj has served as expert for the ERC, the DFG (Germany), the ISF (Israel) and the ANR.

11.1.6 Research administration

Since 2020, Christophe Biernacki joined the scientific head of Inria at the national level by acting as a deputy scientific direction in charge of the domain “Applied mathematics, computation and simulation".

Benjamin Guedj is an elected member of the board of the Evaluation Committee.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Hemant Tyagi is teaching
  - Master: Statistics I, 24h, M1, Centrale Lille, France (Nov. 2021 - 13 Dec. 2021)
- 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 (on maternity leave between January and July 2021) taught between September and December 2021
  - Master: Biostatistics, 22.5h, M1, Université de Lille (Faculty of Medicine), France
  - Master: Supervised classification, 20h, M1, Polytech’Lille, France
  - Master: Biostatistics, 35h, M1, Université de Lille (Departments of Computer Science and Biology), France
  - Doctorat: Introduction to statistical analysis of omic data, 12.5h, Université de Lille (Faculty of Medicine), France
  - Doctorat: Statistical analysis of RNA-Seq data, 12.5h, 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

- Christophe Biernacki is teaching four lessons in statistics with missing data for the "Ateliers statistiques de la SFdS".
- 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:

- Eglantine Karle, November 2020, Hemant Tyagi and Cristian Preda
- Guillaume Braun, January 2020, Christophe Biernacki and Hemant Tyagi
- Wilfried Heyse, 2019, Guillemette Marot and Vincent Vandewalle
- Axel Potier, Sale prediction for low turn-over products, November 2020, Christophe Biernacki, Matthieu Marbac, Vincent Vandewalle
- Etienne Kronert, Détection d'anomalie à noyau reproduisant appliquée au domaine IT, Septembre 2020, Alain Celisse et Cristian Preda.
- Issam Moindje, Analyse de données fonctionnelles pour l'identification des biomarqueurs dans l'EEG et le MEG chez les prématurés et les foetus, October 2020, Sophie Dabo, Cristian Preda.
- Luxin Zhang, Model Agnostic Domain Adaptation: application to Fraud Detection, February 2019, Christophe Biernacki, Pascal Germain, Yacine Kecassi
- Filippo Antonazzo, Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach, October 2019, Christophe Biernacki, Christine Keribin
- Clarisse Boinay, anomaly detection and change point detection in contextual dynamic asynchronous graphs with applications in OT cybersecurity, December 2021, Christophe Biernacki, Cristian Preda
- Felix Biggs, PAC-Bayes, deep neural networks and generative models. Started Sept 2019, University College London, supervisors Benjamin Guedj and John Shawe-Taylor.
- Antoine Vendeville, Graph models for cybersecurity and information diffusion on networks. Started Sept 2019, University College London, supervisors Benjamin Guedj and Shi Zhou.
- Antonin Schrab, PAC-Bayes, generative models and hypothesis testing. Started Sept 2020, University College London, supervisors Benjamin Guedj and Arthur Gretton.
- Reuben Adams, PAC-Bayes theory and computational statistics. Started Sept 2020, University College London, supervisors Benjamin Guedj and John Shawe-Taylor.
- Maxime Haddouche, PAC-Bayes, representation learning and online learning. Started Sept 2021, University College London, supervisors Benjamin Guedj and John Shawe-Taylor.
- Théophile Cantelobre, PAC-Bayes, kernel methods and representation learning. Started Sept 2021, University College London, supervisors Benjamin Guedj, Alessandro Rudi and Carlo Ciliberto.

- Mathieu Alain, PAC-Bayes and information theory. Started Sept 2021, University College London, supervisors Benjamin Guedj and Miguel Rodrigues.
- Antoine Picard, Agrégation d'experts et apprentissage multi-tâches : application à la modélisation du processus de méthanisation pour l'optimisation de la gestion de déchêts organiques. Started Sept 2021, CIFRE Suez, supervisors Benjamin Guedj, Roman Moscoviz and Gilles Faÿ.

11.2.3 Juries

- Guillemette Marot acted as an examinator for the PhD thesis of Nathanaël Randriamihamison, Oct 2021 (Univ Toulouse)
- Christophe Biernacki acted as a reviewer for two PhD theses and as an examinator for two PhD theses
- Cristian Preda acted as a referee for the PHD thesis of Zaineb Smida, Université de Monpellier 2, Décembre 2021.
- Cristian Preda acted as a referee for the PHD thesis of Mohamed Es. Sebaiy, Université Cadi Ayyad, Marrakesh , Maroc, Mars 2021.

11.3 Popularization

11.3.1 Interventions

Christophe Biernacki gave a talk to "Session Olympique Universitaire" in September 2021 [37]

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] F. Biggs and B. Guedj. 'Differentiable PAC-Bayes Objectives with Partially Aggregated Neural Networks'. In: *Entropy* (2021). DOI: 10.3390/e23101280. URL: https://hal.inria.fr/hal-02879216.

[12] M. Cucuringu, A. V. Singh, D. Sulem and H. Tyagi. 'Regularized spectral methods for clustering signed networks'. In: *Journal of Machine Learning Research* 22.264 (Nov. 2021), pp. 1–79. URL: https://hal.inria.fr/hal-03101710.

[13] M. Cucuringu and H. Tyagi. 'An extension of the angular synchronization problem to the heterogeneous setting'. In: *Foundations of Data Science* (2022). URL: https://hal.inria.fr/hal-03101682.

[14] A. D'Aspremont, M. Cucuringu and H. Tyagi. 'Ranking and synchronization from pairwise measurements via SVD'. In: *Journal of Machine Learning Research* 22.19 (11th Feb. 2021), pp. 1–63. URL: https://hal.archives-ouvertes.fr/hal-02340372.

[15] D. Duca, C. Pirim, M. Vojkovic, Y. Carpentier, A. Faccinetto, M. Ziskind, C. Preda and C. Focsa. 'A novel laser-based method to measure the adsorption energy on carbonaceous surfaces'. In: *Carbon* 173 (Mar. 2021), pp. 540–556. DOI: 10.1016/j.carbon.2020.10.064. URL: https://hal.archives-ouvertes.fr/hal-03141569.

[16] A. Ehrhardt, C. Biernacki, V. Vandewalle, P. Heinrich and S. Beben. 'Reject Inference Methods in Credit Scoring'. In: *Journal of Applied Statistics* (23rd Feb. 2021). URL: https://hal.inria.fr/hal-03087279.

[17] M. Fanuel and H. Tyagi. 'Denoising modulo samples: k-NN regression and tightness of SDP relaxation'. In: *Information and Inference* (13th Oct. 2021). URL: https://hal.inria.fr/hal-03101740.

[18] A. Gbogbo, B. Kouakou, S. Dabo-Niang and J. Zoueu. 'Predictive model for airborne insect abundance intercepted by a continuous wave Scheimpflug lidar in relation to meteorological parameters'. In: *Ecological Informatics* 68 (May 2022), p. 101528. DOI: 10.1016/j.ecoinf.2021.101528. URL: https://hal.inria.fr/hal-03527459.

[19] B. Guedj and L. Li. 'Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly'. In: *Entropy* (2021). DOI: 10.3390/e23111534. URL: https://hal.inria.fr/hal-01796011.

[20] B. Guedj and L. Pujol. 'Still no free lunches: the price to pay for tighter PAC-Bayes bounds'. In: *Entropy* (2021). DOI: 10.3390/e23111529. URL: https://hal.inria.fr/hal-02401286.

[21] M. Haddouche, B. Guedj, O. Rivasplata and J. Shawe-Taylor. 'PAC-Bayes unleashed: generalisation bounds with unbounded losses'. In: *Entropy* (2021). DOI: 10.3390/e23101330. URL: https://hal.inria.fr/hal-02872173.

[22] S.-H. A. Kanga, O. Hili and S. Dabo-Niang. 'On Nonparametric Conditional Quantile Estimation for Non-stationary Random'. In: *Afrika Statistika* 16.4 (1st Oct. 2021), pp. 3009–3039. DOI: 10.16929/as/2021.3009.193. URL: https://hal.inria.fr/hal-03527760.

[23] A. S. Khelifa, C. Guillen Sanchez, K. M. Lesage, L. Huot, T. Mouveaux, P. Péricard, N. Barois, H. Touzet, G. Marot, E. Roger and M. Gissot. 'TgAP2IX-5 is a key transcriptional regulator of the asexual cell cycle division in Toxoplasma gondii'. In: *Nature Communications* 12 (7th Jan. 2021), p. 116. DOI: 10.1038/s41467-020-20216-x. URL: https://hal.archives-ouvertes.fr/hal-03107529.

[24] 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.

[25] 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.

[26] M. Selosse, J. Jacques and C. Biernacki. 'ordinalClust: An R Package to Analyze Ordinal Data'. In: *The R Journal* 12.2 (14th Jan. 2021). DOI: 10.32614/RJ-2021-011. URL: https://hal.inria.fr/hal-01678800.

[27] H. Tyagi. 'Error analysis for denoising smooth modulo signals on a graph'. In: *Applied and Computational Harmonic Analysis* (7th Dec. 2021). URL: https://hal.inria.fr/hal-03101720.

[28] A. Vendeville, B. Guedj and S. Zhou. 'Forecasting elections results via the voter model with stubborn nodes'. In: *Applied Network Science* (7th Jan. 2021). URL: https://hal.inria.fr/hal-02946434.

International peer-reviewed conferences

[29] V. Cohen-Addad, B. Guedj, V. Kanade and G. Rom. 'Online k-means Clustering'. In: AISTATS 2021 - The 24th International Conference on Artificial Intelligence and Statistics. Virtual, France, 2021. URL: https://hal.inria.fr/hal-02401290.

[30] W. Heyse, V. Vandewalle, P. Amouyel, G. Marot, C. Bauters and F. Pinet. 'Identification of patients subtypes based on protein expression for prediction of heart failure after myocardial infarction'. In: Printemps de la cardiologie 2021. Vol. 13. 2. Online, France, May 2021, p. 213. DOI: 10.1016/j.acvdsp.2021.04.159. URL: https://hal.inria.fr/hal-03525354.

[31] A. Vendeville, B. Guedj and S. Zhou. 'Towards control of opinion diversity by introducing zealots into a polarised social group'. In: Complex Networks and Their Applications X. Madrid, Spain, 30th Nov. 2021. DOI: 10.1007/978-3-030-93413-2_29. URL: https://hal.inria.fr/hal-02872161.

[32] V. Zantedeschi, M. J. Kusner and V. Niculae. 'Learning Binary Decision Trees by Argmin Differentiation'. In: International Conference on Machine Learning. Virtual, United Kingdom, 7th June 2021. URL: https://hal.inria.fr/hal-03399069.

[33] 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: NeurIPS. Online, France, 2021. URL: https://hal.inria.fr/hal-03278470.

National peer-reviewed Conferences

[34] M. Marbac, M. Sedki, C. Biernacki and V. Vandewalle. 'Simultaneous semi-parametric estimation of clustering and regression'. In: 52èmes journées de la SFdS. Nice / Virtual, France, 7th June 2021. URL: https://hal.inria.fr/hal-03515286.

[35] E. Nyangwile, W. Heyse, C. Méjean and J. Dallongeville. 'Analyses des transitions alimentaires dans le monde entre 1961 à 2018'. In: Journées Francophones de Nutrition (JFN 2021). 2021. Online, France: JFN; 2021-11-12, 2021. URL: https://hal.inrae.fr/hal-03463798.

Conferences without proceedings

[36] F. Antonazzo, C. Biernacki and C. Keribin. 'Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach'. In: Working Group - Model-based Clustering. Athens, Greece, 25th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505673.

[37] C. Biernacki. 'Le sport face aux jeux (de données)'. In: Session Olympique Universitaire (SOU). Lille, France, 24th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505657.

[38] C. Biernacki, C. Boyer, G. Celeux, J. Josse, F. Laporte, M. M. Lourdelle and A. Sportisse. 'Dealing with missing data in model-based clustering through a MNAR model'. In: The 14th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Zakopane, Poland, 11th May 2021. URL: https://hal.archives-ouvertes.fr/hal-03505659.

[39] C. Biernacki, C. Boyer, G. Celeux, J. Josse, F. Laporte, M. Marbac Lourdelle, A. Sportisse and V. Vandewalle. 'Impact of Missing Data on Mixtures and Clustering'. In: MHC2021 - Mixtures, Hidden Markov Models, Clustering. Orsay, France, 2nd June 2021. URL: https://hal.archives-ouvertes.fr/hal-03505664.

[40] C. Biernacki, M. Marbac Lourdelle and V. Vandewalle. 'Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering'. In: MIMO 2021: Workshop on Mixture Models. Rouen, France, 8th Apr. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505667.

[41] G. Braun, H. Tyagi and C. Biernacki. 'Clustering multilayer graphs with missing nodes'. In: The 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021). « Virtual », France, 13th Apr. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505636.

[42] W. Heyse, V. Vandewalle, P. Amouyel, G. Marot, C. Bauters and F. Pinet. 'Support of temporal structure in the statistical analysis of high-throughput proteomic data'. In: Journées de Statistique 2021. Nice, France, 7th June 2021. URL: https://hal.inria.fr/hal-03525345.

[43] M. Perez-Ortiz, O. Rivasplata, E. Parrado-Hernandez, B. Guedj and J. Shawe-Taylor. 'Progress in Self-Certified Neural Networks'. In: NeurIPS 2021 - Conference on Neural Information Processing Systems. Session Workshop : Bayesian Deep Learning. Virtual, United Kingdom, 6th Dec. 2021. URL: https://hal.inria.fr/hal-03430821.

[44] C. S. de Witt, C. Tong, V. Zantedeschi, D. De Martini, F. Kalaitzis, M. Chantry, D. Watson-Parris and P. Bilinski. 'RainBench: Towards Global Precipitation Forecasting from Satellite Imagery'. In: Association for the Advancement of Artificial Intelligence. Virtual, United Kingdom, 1st Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03425405.

[45] 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.archives-ouvertes.fr/hal-03505639.

Scientific book chapters

[46] M. S. Ahmed, L. Broze, S. Dabo-Niang and Z. Gharbi. 'Functional Linear Spatial Autoregressive Models'. In: *Geostatistical Functional Data Analysis: Theory and Methods*. Wiley, 2021. URL: https://hal.archives-ouvertes.fr/hal-01810819.

[47] F. Antonazzo, C. Biernacki and C. Keribin. 'A binned technique for scalable model-based clustering on huge datasets'. In: *Book of Short Papers of the 5th international workshop on Models and Learning for Clustering and Classification MBC2 2020, Catania, Italy*. 26th Oct. 2021, pp. 11–16. URL: https://hal.archives-ouvertes.fr/hal-03097284.

[48] S. Dabo-Niang, C. Ternynck, B. Thiam and A.-F. Yao. 'Non-parametric statistical analysis of spatially distributed functional data'. In: *Geostatistical Functional Data Analysis*. Wiley, 2021. URL: https://hal.archives-ouvertes.fr/hal-01812238.

[49] V. Vandewalle, C. Preda and S. Dabo-Niang. 'Clustering spatial functional data'. In: *Geostatistical Functional Data Analysis : Theory and Methods. Editors: Jorge Mateu, Ramon Giraldo*. Geostatistical Functional Data Analysis : Theory and Methods. John Wiley and Sons, Chichester. ISBN : 978-1-119-38784-8, 1st Jan. 2021. URL: https://hal.inria.fr/hal-01948934.

Doctoral dissertations and habilitation theses

[50] G. Marot. 'Contributions méthodologiques en statistique pour l'analyse et l'intégration de données -omiques et cliniques'. Université de Lille, 8th Jan. 2021. URL: https://hal.archives-ouvertes.fr/tel-03440914.

[51] V. Vandewalle. 'Contribution to model-based clustering of heterogeneous data'. Université de Lille, 7th Jan. 2021. URL: https://hal.inria.fr/tel-03118189.

Reports & preprints

[52] C. Agonkoui, S. Dabo-Niang and F. Djibril Moussa. *Multivariate Functional Principal Component Analysis for stratified data*. 15th Jan. 2022. URL: https://hal.inria.fr/hal-03527475.

[53] 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.

[54] F. Antonazzo, C. Biernacki and C. Keribin. *Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach*. 17th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03485364.

[55] F. Biggs and B. Guedj. *Non-Vacuous Generalisation Bounds for Shallow Neural Networks*. 4th Feb. 2022. URL: https://hal.inria.fr/hal-03557415.

[56] F. Biggs and B. Guedj. *On Margins and Derandomisation in PAC-Bayes*. 9th July 2021. URL: https://hal.inria.fr/hal-03282597.

[57] G. Braun, H. Tyagi and C. Biernacki. *An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees*. 14th Jan. 2022. URL: https://hal.inria.fr/hal-03526257.

[58] D. Collard, S. Dabo-Niang and M. Tarhan. *Single cell classification using statistical learning on mechanical properties measured by mems tweezers*. 17th Jan. 2022. URL: https://hal.inria.fr/hal-03528082.

[59] S. Dabo-Niang, S. Doumun and J. T. Zoueu. *A Novel Unstained Blood Smears Multispectral Images Normalization. Application to Unstained Malaria Infected Blood Smear*. 7th Feb. 2021. URL: https://hal.inria.fr/hal-03133828.

[60] S. Dabo-Niang, D. Pathmanathan and A. A. Hassan. *Functional spatial principal Component Analysis and Application to demography*. 7th Feb. 2021. URL: https://hal.inria.fr/hal-03133827.

[61] S. Dabo-Niang, D. Pathmanathan and H. Omar. *Clustering DNA sequences for phylogenetic trees using a functional data framework*. 7th Feb. 2021. URL: https://hal.inria.fr/hal-03133825.

[62] 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.

[63] M. Fanuel and H. Tyagi. *Recovering Hölder smooth functions from noisy modulo samples*. 4th Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03511325.

[64] 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.

[65] E. Karlé and H. Tyagi. *Dynamic Ranking with the BTL Model: A Nearest Neighbor based Rank Centrality Method*. 10th Jan. 2022. URL: https://hal.inria.fr/hal-03519271.

[66] S. Nasini, T. R. Tchouya and S. Dabo-Niang. *An estimation framework for the influential-imitator diffusion*. 15th Jan. 2022. URL: https://hal.inria.fr/hal-03527489.

[67] M. Pérez-Ortiz, O. Rivasplata, B. Guedj, M. Gleeson, J. Zhang, J. Shawe-Taylor, M. Bober and J. Kittler. *Learning PAC-Bayes Priors for Probabilistic Neural Networks*. 22nd Sept. 2021. URL: https://hal.inria.fr/hal-03351794.

[68] F. Sainul, S. Dabo-Niang and A. Adekpedjou. *Modelling Spatially Clustered Failure Time Data via Multivariate Gaussian Random Fields*. 15th Jan. 2022. URL: https://hal.inria.fr/hal-03527477.

[69] A. Schrab, B. Guedj and A. Gretton. *KSD Aggregated Goodness-of-fit Test*. 3rd Feb. 2022. URL: https://hal.inria.fr/hal-03554423.

[70] A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj and A. Gretton. *MMD Aggregated Two-Sample Test*. 29th Oct. 2021. URL: https://hal.inria.fr/hal-03408976.

[71] A. Sportisse, C. Biernacki, C. Boyer, J. Josse, M. Marbac Lourdelle, G. Celeux and F. Laporte. *Model-based Clustering with Missing Not At Random Data*. 17th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03494674.

[72] T. Tchamie, S. Dabo-Niang and A. Diop. *Estimation of extreme tail index for β–mixing random fields*. 7th Feb. 2021. URL: https://hal.inria.fr/hal-03133830.

[73] T. Tchazino, S. Dabo-Niang and A. Diop. *Tail and quantile estimation for real-valued β-mixing spatial data*. 30th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03457114.

[74] T. R. Tchouya, S. Nasini and S. Dabo-Niang. *An asymptotic approximation for the extended Bass diffusion model and application to pandemic outbreaks*. 7th Feb. 2021. URL: https://hal.inria.fr/hal-03133829.

[75] L. Zhang, P. Germain, Y. Kessaci and C. Biernacki. *Interpretable Domain Adaptation Using Unsupervised Feature Selection on Pre-trained Source Models*. 24th Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03325509.

Other scientific publications

[76] B. Guedj. *Covid-19 and AI: Unexpected Challenges and Lessons*. 7th July 2021. URL: https://hal.inria.fr/hal-03277494.

[77] C. Preda, Q. Grimonprez and V. Vandewalle. *Categorical functional data analysis. The cfda R package*. London, United Kingdom, 18th Dec. 2021. URL: https://hal.inria.fr/hal-03518940.

[78] C. Preda, Q. Grimonprez and V. Vandewalle. *Categorical functional data analysis.The cfda R package*. Nice, France, 7th June 2021. URL: https://hal.inria.fr/hal-03519016.

12.3 Other

Scientific popularization

[79] S. Boumeddane, L. Hamdad, H. Haddadou and S. Dabo-Niang. 'Dimensionality Reduction and Bandwidth Selection for Spatial Kernel Discriminant Analysis'. In: 13th International Conference on Agents and Artificial Intelligence. Online Streaming, New Zealand: SCITEPRESS - Science and Technology Publications, 4th Feb. 2021, pp. 278–285. DOI: 10.5220/0010269002780285. URL: https://hal.inria.fr/hal-03527454.

Educational activities

[80] F. Antonazzo, C. Biernacki and C. Keribin. 'Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach'. Doctoral. France, 29th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505670.

[81] C. Biernacki. 'Traitement statistique des donn´ees manquantes-Part III Missing not at random data (MNAR)'. Doctoral. France, 11th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505651.

[82] C. Biernacki. 'Traitement statistique des donn´ees manquantes-Part IV Binned data for big data analysis'. Doctoral. France, 11th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505653.

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

[84] C. Biernacki. 'Traitement statistique des données manquantes-Part IINumerical and non-numerical data'. Doctoral. France, 10th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505650.

12.4 Cited publications

[85] C. Biernacki, M. Marbac and V. Vandewalle. 'Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering'. In: *Journal of Classification* (July 2020). DOI: 10.1007/s00357-020-09369-y. URL: https://hal.archives-ouvertes.fr/hal-01949155.

[86] F. Dewez, B. Guedj and V. Vandewalle. 'From industry-wide parameters to aircraft-centric on-flight inference: improving aeronautics performance prediction with machine learning'. In: *Data-Centric Engineering* (Oct. 2020). DOI: 10.1017/dce.2020.12. URL: https://hal.inria.fr/hal-02570875.

[87] H. Frydman. 'Estimation in the Mixture of Markov Chains Moving With Different Speeds'. In: *Journal of the American Statistical Association* 100.471 (2005), pp. 1046–1053.

[88] R. Gupta, R. Kumar and S. Vassilvitskii. 'On mixtures of Markov chains'. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Citeseer. 2016, pp. 3449–3457.

[89] V. Vandewalle, A. Caron, C. Delettrez, R. Périchon, S. Pelayo, A. Duhamel and B. Dervaux. 'Estimating the number of usability problems affecting medical devices: modelling the discovery matrix'. In: *BMC Medical Research Methodology* 20.1 (Sept. 2020). DOI: 10.1186/s12874-020-01091-y. URL: https://hal.archives-ouvertes.fr/hal-03117742.