RESEARCH CENTRE

Inria Centre at the University of Lille

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2023

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

A3.1.10. – Heterogeneous data

A3.2.3. – Inference

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.5. – Bayesian methods

A3.4.7. – Kernel methods

A5.2. – Data visualization

A5.9.2. – Estimation, modeling

A6.2.3. – Probabilistic methods

A6.2.4. – Statistical methods

A6.3.3. – Data processing

A9.2. – Machine learning

Other research topics and application domains

B2.2.3. – Cancer

B9.5.6. – Data science

B9.6.3. – Economy, Finance

B9.6.5. – Sociology

1 Team members, visitors, external collaborators

Research Scientists

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

Faculty Members

- Cristian Preda [Team leader, UNIV LILLE, Professor, HDR]
- Sophie Dabo [UNIV LILLE, Professor, HDR]
- Guillemette Marot [UNIV LILLE, Professor, HDR]

Post-Doctoral Fellow

- Rim Essifi [INRIA, Post-Doctoral Fellow, until Aug 2023]

PhD Students

- Reuben Adams [UCL]
- François Bassac [Decathlon, CIFRE]
- Clarisse Boinay [Seckiot]
- Violaine Courrier [WITHINGS, from Sep 2023]
- Clara Dubois [LABO TIMC, from Jun 2023]
- Maxime Haddouche [UNIV LILLE]
- Wilfried Heyse [UNIV LILLE, until Aug 2023]
- Eglantine Karle [INRIA, until Oct 2023]
- Etienne Kronert [WORLDLINE]
- Issam Ali Moindjie [INRIA]
- Axel Potier [ADEO]
- Antonin Schrab [UCL]

Technical Staff

- Ernesto Javier Araya Valdivia [INRIA, Engineer, from Mar 2023 until Oct 2023]
- Rachid Boulkhir [INRIA, Engineer, until Sep 2023]
- Guillaume Braun [INSEE, until Mar 2023]
- Ismat Yahia Chaib Draa [ALICANTE, Engineer, until Nov 2023]
- Louise Chen [INRIA, Engineer, from Nov 2023]

Interns and Apprentices

- Paguidame Sambiani [INRIA, Intern, from Jul 2023 until Sep 2023]

Administrative Assistant

- Anne Rejl [INRIA]

External Collaborator

- Alain Celisse [UNIV PARIS I, HDR]

2 Overall objectives

2.1 Context

In several respects, modern society has strengthened the need for statistical analysis both from the applied and theoretical points of view. The genesis comes from the easier availability of data thanks to technological breakthroughs (storage, transfer, computing), and are now so widespread that they are no longer limited to large human organizations. The more or less conscious goal of such data availability is the expectation of improving the quality of "since the dawn of time" statistical stories which are namely discovering new knowledge or doing better predictions. These both central tasks can be referred to respectively as unsupervised learning or supervised learning, even if it is not limited to them or other names exist depending on communities. Somehow, it pursues the following hope: "more data for better quality and more numerous results".

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

2.2 Goals

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

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

3 Research program

3.1 Research axis 1: Unsupervised learning

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

3.2 Research axis 2: Performance assessment

One main concern of the Modal team is to provide theoretical justifications on the procedures which are designed. Such guarantees are important to avoid misleading conclusions resulting from any unsuitable use. For example, one ingredient in proving these guarantees is the use of the PAC framework, leading to finite-sample concentration inequalities. More precisely, contributions to PAC learning rely on the classical empirical process theory and the PAC-Bayesian theory. The Modal team exploits such non-asymptotic tools to analyze the performance of iterative algorithms (such as gradient descent), cross-validation estimators, online change-point detection procedures, ranking algorithms, matrix factorization techniques and clustering methods, for instance. The team also develops some expertise on the formal dynamic study of algorithms related to mixture models (important models used in the previous unsupervised setting), like degeneracy for EM algorithm or also label switching for Gibbs algorithm.

3.3 Research axis 3: Functional data

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

3.4 Research axis 4: Applications motivating research

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

4 Application domains

4.1 Economic world

The Modal team applies its research to the economic world through CIFRE PhD supervision such as CACF (credit scoring), A-Volute (expert in 3D sound), Meilleur Taux (insurance comparator), Worldline. It also has several contracts with companies such as COLAS, Nokia-Apsys/Airbus, Safety Line (through the PERF-AI consortium), Agence d'Urbanisme Métropole Européenne de Lille, ASYGN SAS (MEMs, joint Cytomems ANR project), HORIBA France SAS (Raman spectrometry), Withings (medical devices), Seckiot (cyber-security).

4.2 Biology and health

The second main application domain of the team is biology and health. Some members of the team are involved in the direction of Bilille, the bioinformatics platform of Lille, and of OncoLille Institute. Some members of the team also co-supervise PhD students of Inserm teams.

5 Social and environmental responsibility

MODAL has not any social and environmental responsibility.

6 New software, platforms, open data

6.1 New software

6.1.1 MixtComp.V4

Keywords: Clustering, Statistics, Missing data, Mixed data

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

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

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

Contact: Christophe Biernacki

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

Partners: Université de Lille, CNRS

6.1.2 cfda

Name: Categorical functional data analysis

Keyword: Functional data

Functional Description: The R package cfda performs:

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

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

Contact: Cristian Preda

Participants: Cristian Preda, Quentin Grimonprez, Vincent Vandewalle

Partner: Université de Lille

6.1.3 ClusPred

Name: Simultaneous Semi-Parametric Estimation of Clustering and Regression

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

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

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

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

Contact: Matthieu Marbac-Lourdelle

6.1.4 visCorVar

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

Keywords: Data integration, Visualization

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

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

Contact: Guillemette Marot

Participants: Maxime Brunin, Guillemette Marot, Pierre Pericard

Partner: Université de Lille

6.1.5 metaRNASeq

Name: RNA-Seq data meta-analysis

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

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

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

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

Contact: Guillemette Marot

Participants: Guillemette Marot, Andrea Rau, Samuel Blanck

Partners: INRAE, Université de Lille

6.1.6 HDSpatialScan

Name: Multivariate and Functional Spatial Scan Statistics

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

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

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

Contact: Sophie Dabo

6.1.7 MLGL

Name: Multi-Layer Group Lasso

Keywords: Variable selection, Statistical learning

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

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

Contact: Guillemette Marot

6.2 New platforms

6.2.1 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.

7 New results

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

Participants: Christophe Biernacki.

We advocate that co-clustering, is of particular interest to perform high dimension (HD) clustering of individuals even if it is not its primary mission. Indeed, column clustering is recast as a strategy to control the variance of the estimation, the model dimension being driven by the number of groups of variables instead of the number of variables itself. A survey paper published in an international journal [14] advocates the ability of co-clustering to outperform simple mixture row-clustering, even if co-clustering clearly corresponds to a misspecified model situation, revealing a promising manner to efficiently address (very) HD clustering.

It is a joint work with Julien Jacques from University Lyon 2 and Christine Keribin from University Paris-Saclay.

7.2 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.

A preprint has been submited to an international journal [57] and a invited talk to an international conference has been given related to this topic [35].

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.

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

Participants: Christophe Biernacki.

A generic method is introduced to visualize in a Gaussian-like way, and onto $\mathbb{R}^2$ 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 previously in an international journal but has been presented this year as an invited plenary sesssion to the French classification conference [36] and also as an invited talk to the main international workshop on model-based clustering [42].

This is a joint work with Matthieu Marbac from ENSAI and Vincent Vandewalle from University Côte d'Azur.

7.4 Axis 1: Levels Merging in the Latent Class Model

Participants: Christophe Biernacki.

The latent class model (LCM), dedicated to cluster categorical variables, suffers for the curse of dimension when the number of levels is large, situation frequently encountered in practice. We propose to extent LCM to a natural modeling which limits the number of levels by merging them, process which is also equivalent to a specific levels clustering. Related estimation and model selection processes are also presented and discussed. This work has been presented for an invited talk at two international conferences [32], [34].

7.5 Axis 1: Comparative study of series clustering models multivariate temporal data from connected medical objects

Participants: Christophe Biernacki, Violaine Courrier, Cristian Preda.

In healthcare, patient data are often collected in the form of multivariate time series, providing a comprehensive overview of a patient's health status over time. These data are generally scattered and episodic. However, connected medical objects can increase the frequency of data. The objective is to create unsupervised patient profiles from these time series. In the absence of labels, a predictive model can be used to predict future values while performing a space of latent clusters, evaluated according to predictive performance. Using real data from the Withings company, we compare the static clustering approach MAGMACLUST, which creates a cluster at the scale of the entire time series, and the dynamic clustering DGM2, which allows an individual's membership in a group to change over time. This work will be presented to a conference in 2024 [41].

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

Participants: Eglantine Karle, 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 tournament – the 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 $(G_{t'})_{t'\in\mathcal{T}}$ on a regular grid $\mathcal{T}\subset[0,1]$, we aim at recovering the latent strengths of the items $w_t^*\in\mathbb{R}^n$ 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 $(G_{t'})_{t'\in\mathcal{T}}$ 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 $|\mathcal{T}|$. We also complement our theoretical analysis with experiments on real and synthetic data. This work appeared in the Journal of Machine Learning Research [22].

7.7 Axis 1&2: Dynamic Ranking and Translation Synchronization

Participants: Ernesto Araya, Eglantine Karle, Hemant Tyagi.

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

This work appeared in the journal Information and Inference: a journal of the IMA [13].

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

Participants: Guillaume Braun, Hemant Tyagi.

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

This work appeared in the journal Information and Inference: a journal of the IMA [16].

7.9 Axis 1&2: Graph Matching via convex relaxation to the simplex

Participants: Ernesto Araya, Hemant Tyagi.

This paper addresses the Graph Matching problem, which consists of finding the best possible alignment between two input graphs, and has many applications in computer vision, network deanonymization and protein alignment. A common approach to tackle this problem is through convex relaxations of the NP-hard *Quadratic Assignment Problem* (QAP). Here, we introduce a new convex relaxation onto the unit simplex and develop an efficient mirror descent scheme with closed-form iterations for solving this problem. Under the correlated Gaussian Wigner model, we show that the simplex relaxation admits a unique solution with high probability. In the noiseless case, this is shown to imply exact recovery of the ground truth permutation. Additionally, we establish a novel sufficiency condition for the input matrix in standard greedy rounding methods, which is less restrictive than the commonly used ‘diagonal dominance’ condition. We use this condition to show exact one-step recovery of the ground truth (holding almost surely) via the mirror descent scheme, in the noiseless setting. We also use this condition to obtain significantly improved conditions for the GRAMPA algorithm [Fan et al. 2019] in the noiseless setting. Our method is evaluated on both synthetic and real data, demonstrating superior statistical performance compared to existing convex relaxation methods with similar computational costs.

This work is currently under review in a journal [60].

7.10 Axis 2: Learning linear dynamical systems under convex constraints

Participants: Hemant Tyagi.

We consider the problem of finite-time identification of linear dynamical systems from T samples of a single trajectory. Recent results have predominantly focused on the setup where no structural assumption is made on the system matrix $A^* \in R^{n \times n}$, and have consequently analyzed the ordinary least squares (OLS) estimator in detail. We assume prior structural information on A^* is available, which can be captured in the form of a convex set $\mathcal{K}$ containing A^*. For the solution of the ensuing constrained least squares estimator, we derive non-asymptotic error bounds in the Frobenius norm that depend on the local size of $\mathcal{K}$ at A^*. To illustrate the usefulness of these results, we instantiate them for three examples, namely when (i) A^* is sparse and $\mathcal{K}$ is a suitably scaled ℓ_1 ball; (ii) $\mathcal{K}$ is a subspace; (iii) $\mathcal{K}$ consists of matrices each of which is formed by sampling a bivariate convex function on a uniform $n \times n$ grid (convex regression). In all these situations, we show that A^* can be reliably estimated for values of T much smaller than what is needed for the unconstrained setting.

This work is currently under review in a journal [59] and is joint work with Denis Efimov (Inria Lille, Valse team).

7.11 Axis 2: An estimation approach for the influential–imitator diffusion

Participants: Sophie Dabo-Niang.

This paper presents a numerical estimation procedure for the influential–imitator diffusion, an extension to the Bass model in which a population is partitioned into two segments: influentials (who influence each other) and imitators (whose choices are affected by the ones of influentials). Focusing on the estimation of the model parameters, we propose a maximum likelihood approach and investigate its numerical solvability, building on an asymptotic approximation of the underlying differential equation. Specifically, we develop a truncated series expansion, exhibiting an increasing accuracy when the spontaneous innovation decreases. After uncovering the theoretical properties of the proposed methodology, we propose a specialized block coordinate descent method for the numerical maximization of the likelihood function. Empirical and computational tests are provided using the Michell and West dataset about the cannabis consumption of a cohort of students over their second, third and fourth year at a secondary school in Glasgow. The estimated imitation pattern confirms the well-known hypothesis on peer influences, where the choices of popular children represent the leading effects to determine the habits of others.

It is a joint work with Ringo Thomas Tchouya (IMSP, bENIN), Stefano Nasini (IESEG, Lille) [31].

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

Participants: Sophie Dabo-Niang.

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

It is a joint work with Mohamed Salem Ahmed (University of Lille, CERIM), Mohamed Attouch (University Sidi Bel Abbes, Algeria), Mamadou Ndiaye (UCAD, Senegal) [11].

7.13 Axis 2: FDR control for Online Anomaly Detection

Participants: Etienne Kronert, Alain Célisse, Dalila Hattab.

The goal of anomaly detection is to identify observations generated by a process that is different from a reference one. An accurate anomaly detector must ensure low false positive and false negative rates. However in the online context such a constraint remains highly challenging due to the usual lack of control of the False Discovery Rate (FDR). In particular the online framework makes it impossible to use classical multiple testing approaches such as the Benjamini-Hochberg (BH) procedure. Our strategy overcomes this difficulty by exploiting a local control of the "modified FDR" (mFDR). An important ingredient in this control is the cardinality of the calibration set used for computing empirical p-values, which turns out to be an influential parameter. It results a new strategy for tuning this parameter, which yields the desired FDR control over the whole time series. The statistical performance of this strategy is analyzed by theoretical guarantees and its practical behavior is assessed by simulation experiments which support our conclusions. See for more details [54].

7.14 Axis 2: Optimistic Dynamic Regret Bounds

Participants: Benjamin Guedj, Maxime Haddouche.

Online Learning (OL) algorithms have originally been developed to guarantee good performances when comparing their output to the best fixed strategy. The question of performance with respect to dynamic strategies remains an active research topic. We develop in this work dynamic adaptations of classical OL algorithms based on the use of experts' advice and the notion of optimism. We also propose a constructivist method to generate those advices and eventually provide both theoretical and experimental guarantees for our procedures.

Joint work with Olivier Wintenberger (Sorbonne Université). See for more details [49].

7.15 Axis 2: Wasserstein PAC-Bayes Learning: Exploiting Optimisation Guarantees to Explain Generalisation

Participants: Benjamin Guedj, Maxime Haddouche.

PAC-Bayes learning is an established framework to both assess the generalisation ability of learning algorithms, and design new learning algorithm by exploiting generalisation bounds as training objectives. Most of the exisiting bounds involve a *Kullback-Leibler* (KL) divergence, which fails to capture the geometric properties of the loss function which are often useful in optimisation. We address this by extending the emerging *Wasserstein PAC-Bayes* theory. We develop new PAC-Bayes bounds with Wasserstein distances replacing the usual KL, and demonstrate that sound optimisation guarantees translate to good generalisation abilities. In particular we provide generalisation bounds for the *Bures-Wasserstein SGD* by exploiting its optimisation properties. See for details [48].

7.16 Axis 2: Generalization Bounds: Perspectives from Information Theory and PAC-Bayes

Participants: Benjamin Guedj.

A fundamental question in theoretical machine learning is generalization. Over the past decades, the PAC-Bayesian approach has been established as a flexible framework to address the generalization capabilities of machine learning algorithms, and design new ones. Recently, it has garnered increased interest due to its potential applicability for a variety of learning algorithms, including deep neural networks. In parallel, an information-theoretic view of generalization has developed, wherein the relation between generalization and various information measures has been established. This framework is intimately connected to the PAC-Bayesian approach, and a number of results have been independently discovered in both strands. In this monograph, we highlight this strong connection and present a unified treatment of generalization. We present techniques and results that the two perspectives have in common, and discuss the approaches and interpretations that differ. In particular, we demonstrate how many proofs in the area share a modular structure, through which the underlying ideas can be intuited. We pay special attention to the conditional mutual information (CMI) framework; analytical studies of the information complexity of learning algorithms; and the application of the proposed methods to deep learning. This monograph is intended to provide a comprehensive introduction to information-theoretic generalization bounds and their connection to PAC-Bayes, serving as a foundation from which the most recent developments are accessible. It is aimed broadly towards researchers with an interest in generalization and theoretical machine learning.

Joint work with Fredrik Hellström (UCL), Giuseppe Durisi (Chalmers), Maxim Raginsky (University of Illinois). See for details [50].

7.17 Axis 2: Comparing Comparators in Generalization Bounds

Participants: Benjamin Guedj.

We derive generic information-theoretic and PAC-Bayesian generalization bounds involving an arbitrary convex comparator function, which measures the discrepancy between the training and population loss. The bounds hold under the assumption that the cumulant-generating function (CGF) of the comparator is upper-bounded by the corresponding CGF within a family of bounding distributions. We show that the tightest possible bound is obtained with the comparator being the convex conjugate of the CGF of the bounding distribution, also known as the Cramér function. This conclusion applies more broadly to generalization bounds with a similar structure. This confirms the near-optimality of known bounds for bounded and sub-Gaussian losses and leads to novel bounds under other bounding distributions.

Joint work with Fredrik Hellström (UCL). See for more details [51].

7.18 Axis 2: Federated Learning with Nonvacuous Generalisation Bounds

Participants: Benjamin Guedj, Maxime Haddouche.

We introduce a novel strategy to train randomised predictors in federated learning, where each node of the network aims at preserving its privacy by releasing a local predictor but keeping secret its training dataset with respect to the other nodes. We then build a global randomised predictor which inherits the properties of the local private predictors in the sense of a PAC-Bayesian generalisation bound. We consider the synchronous case where all nodes share the same training objective (derived from a generalisation bound), and the asynchronous case where each node may have its own personalised training objective. We show through a series of numerical experiments that our approach achieves a comparable predictive performance to that of the batch approach where all datasets are shared across nodes. Moreover the predictors are supported by numerically nonvacuous generalisation bounds while preserving privacy for each node. We explicitly compute the increment on predictive performance and generalisation bounds between batch and federated settings, highlighting the price to pay to preserve privacy.

Joint work with Pierre Jobic (CEA). See for more details [52].

7.19 Axis 2: Learning via Wasserstein-Based High Probability Generalisation Bounds

Participants: Benjamin Guedj, Maxime Haddouche.

Minimising upper bounds on the population risk or the generalisation gap has been widely used in structural risk minimisation (SRM) – this is in particular at the core of PAC-Bayesian learning. Despite its successes and unfailing surge of interest in recent years, a limitation of the PAC-Bayesian framework is that most bounds involve a Kullback-Leibler (KL) divergence term (or its variations), which might exhibit erratic behavior and fail to capture the underlying geometric structure of the learning problem – hence restricting its use in practical applications. As a remedy, recent studies have attempted to replace the KL divergence in the PAC-Bayesian bounds with the Wasserstein distance. Even though these bounds alleviated the aforementioned issues to a certain extent, they either hold in expectation, are for bounded losses, or are nontrivial to minimize in an SRM framework. In this work, we contribute to this line of research and prove novel Wasserstein distance-based PAC-Bayesian generalisation bounds for both batch learning with independent and identically distributed (i.i.d.) data, and online learning with potentially non-i.i.d. data. Contrary to previous art, our bounds are stronger in the sense that (i) they hold with high probability, (ii) they apply to unbounded (potentially heavy-tailed) losses, and (iii) they lead to optimizable training objectives that can be used in SRM. As a result we derive novel Wasserstein-based PAC-Bayesian learning algorithms and we illustrate their empirical advantage on a variety of experiments.

Joint work with Umut Simsekli and Paul Viallard (EP SIERRA, CRI PRO).

See for more details [40].

7.20 Axis 2: A note on regularised NTK dynamics with an application to PAC-Bayesian training

Participants: Benjamin Guedj.

We establish explicit dynamics for neural networks whose training objective has a regularising term that constrains the parameters to remain close to their initial value. This keeps the network in a lazy training regime, where the dynamics can be linearised around the initialisation. The standard neural tangent kernel (NTK) governs the evolution during the training in the infinite-width limit, although the regularisation yields an additional term appears in the differential equation describing the dynamics. This setting provides an appropriate framework to study the evolution of wide networks trained to optimise generalisation objectives such as PAC-Bayes bounds, and hence potentially contribute to a deeper theoretical understanding of such networks.

Joint work with Eugenio Clerico (University of Oxford and Uni Pompeu Fabra).

See for more details [47].

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

Participants: Sophie Dabo-Niang.

In environmental surveillance, cluster detection of environmental black spots is of major interest due to the adverse health effects of pollutants, as well as their known synergistic effect. Thus, this paper introduces three new spatial scan statistics for multivariate functional data, applicable for detecting clusters of abnormal air pollutants concentrations measured spatially at a very fine scale in northern France in October 2021 taking into account their correlations. Mathematically, our methodology is derived from a functional multivariate analysis of variance, an adaptation of the Hotelling T^2-test statistic, and a multivariate extension of the Wilcoxon test statistic. The approaches were evaluated in a simulation study and then applied to the air pollution dataset.

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).

For more details, see [18].

7.22 Axis 3: On estimation and prediction in spatial functional linear regression model

Participants: Sophie Dabo-Niang.

We consider a spatial functional linear regression, where a scalar response is related to a square-integrable spatial functional process. We use a smoothing spline estimator for the functional slope parameter and establish a finite sample bound for variance of this estimator. Then we give the optimal bound of the prediction error under mixing spatial dependence. Finally, we illustrate our results by simulations and by an application to ozone pollution forecasting at nonvisited sites.

It is a joint work with Stéphane Bouka (University of , CERIM), Guy-Martial Nkiet Ahmed (University of France Ville, Gabon), Michaël Genin (University of France Ville, Gabon). For more details, see [15].

7.23 Axis 3: Spatial Autocorrelation of Global Stock Exchanges Using Functional Areal Spatial Principal Component Analysis

Participants: Sophie Dabo-Niang.

This work focuses on functional data presenting spatial dependence. The spatial autocorrelation of stock exchange returns for 71 stock exchanges from 69 countries was investigated using the functional Moran's I statistic, classical principal component analysis (PCA) and functional areal spatial principal component analysis (FASPCA). This work focuses on the period where the 2015–2016 global market sell-off occurred and proved the existence of spatial autocorrelation among the stock exchanges studied. The stock exchange return data were converted into functional data before performing the classical PCA and FASPCA. Results from the Monte Carlo test of the functional Moran's I statistics show that the 2015–2016 global market sell-off had a great impact on the spatial autocorrelation of stock exchanges. Principal components from FASPCA show positive spatial autocorrelation in the stock exchanges. Regional clusters were formed before, after and during the 2015–2016 global market sell-off period. This work explored the existence of positive spatial autocorrelation in global stock exchanges and showed that FASPCA is a useful tool in exploring spatial dependency in complex spatial data.

It is a joint work with Tzung Hsuen Khoo (University of Malaya, Malaysia), Dharini Pathmanathan (University of Malaya, Malaysia).

For more details, see [23].

7.24 Axis 3: Multivariate functional principal component analysis for endogenously stratified data

Participants: Sophie Dabo-Niang.

CWe address the problem of performing dimension reduction on multivariate functional data observed on different domains in an endogenously stratified sampling context. The aim is to propose a new multivariate functional principal component analysis (MFPCA) approach for data sampled by a stratification of a population according to a binary variable of interest. This estimation strategy is derived from a direct relationship between univariate and multivariate FPCA for finite Karhunen-Loève decompositions. The proposed methodology yields encouraging results and can be applied to data with measurement errors. Computational results on simulated data highlight the good performance of the proposed methodology compared to the classical MFPCA, which ignores the type of data sampling. A real-life application considering breast cancer cells data is also presented.

It is a joint work with Idris Christelle Judith Agonkoui (IMSP, Benin), Freedath Djibril Moussa (IMSP, Benin). For more details, see [66].

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

Participants: Issam Moindjie, Sophie Dabo, Cristian Preda.

Multivariate functional data is considered as sample paths of a multivariate valued stochastic process, $X = (X_1, \ldots, X_d)$. In this setting, each dimension X_i, $i = 1, \ldots, d$, is a stochastic process, $X_i = \{X_i(t), t \in \mathscr{I}_i\}$, where $\mathscr{I}_i$ is some compact domaine of $\mathbb{R}$. The problems of linear regression and binary classification are addressed by PLS regularization techniques. For application purposes, decision tree methods combined with functional PLS regression are proposed. For more details, see [25].

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

Participants: Issam Moindjie, Sophie Dabo, Cristian Preda.

Multivariate functional data is considered under the assumption of spatially dependence between dimensions. Each dimension is associated to some (spatial) clusters with potentially different effect on a response variable. In the context of linear regression with multivariate functional data, a natural assumption is to consider the same regression coefficient (slope) function for all dimensions belonging to the same cluster. Fused and group lasso techniques are extended for this purpose. This work was submitted to CSDA journal ([55]) and [45].

7.27 Axis 3: Linear approximation for multivariate categorical functional data

Participants: Cristian Preda, Quentin Grimonprez.

Multivariate categorical functional data can be seen as one-dimensional categorical functional data but with a number of states equal to the product of the number of states for each dimension. That yields to a largy computational complexity that can be avoid by proposing a linear approximation of the optimal encodings. Indeed, the optimal encodings are the conditional expectation of the principal components with respect to functional random vector. In our appproach this conditional espctation is considered as a linear form of the dimensions of the functional vector. See for more details [44] and 6.1.2.

7.28 Axis 4: Multi-layer group Lasso

Participants: Guillemette Marot.

Multi-Layer Group-Lasso (MLGL) is a procedure of variable selection in the context of redundancy between explanatory variables, which holds true with high-dimensional data. The proposed approach combines variable aggregation and selection in order to improve interpretability and performance. The associated R package is available on CRAN and its related publication [19], accepted in 2023 in Journal of Statistical Software, gives more details about the statistical procedure.

7.29 Axis 4: Research of biomarkers using penalised regressions

Participants: Guillemette Marot, Wilfried Heyse.

Thanks to Lasso logistic regression, a joint work with Hélène Sarter identified a novel 8-predictors signature to predict complicated disease course in pediatric-onset Crohn's disease [28]. The research of biomarkers let us the opportunity to test various multi-block approaches in order to combine clinical and omics data. Finally, we retained a very simple approach, which performed the best and offered results that were further validated.

7.30 Axis 4: Research of biomarkers using competing risks models and clustering

Participants: Guillemette Marot, Wilfried Heyse.

When using Lasso penalized Cox regressions on proteomics data to predict heart failure after myocardial infarction, we did not manage to beat predictive models relying on only clinical data. Therefore, we changed the strategy and finally used clustering to identify subtypes of patients based on proteins which could help predict heart failure [21]. The methodology looks simple in the paper but there was a lot of work to define the outcome to keep and to choose the best strategy. The models including only clinical data were already good and it was challenging to obtain better results by including proteomics data. In this project, we experimentated the necessity to take into account the competing risks in the modeling and had to perform univariate analyses for proteomics before multivariate analysis. This work has relied on a strong collaboration with Florence Pinet, specialist in proteomics and Christophe Bauters, cardiologist.

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

Participants: Guillemette Marot.

Our expertise on empirical bayesian approaches for proteomics data analysis has led to a publication in Annals of Rheumatic Diseases (impact factor 28) [27]. This is a joint work with Dr S.Sanges and Pr D. Launay. The proteomic analysis has revealed potential biomarkers that may assist diagnosis and treatment of patients with systemic clerosis-associated pulmonary arterial hypertension (SSc-PAH). Further biological validation in an independent cohort revealed that chemerin, which was highlighted in the exploratory analysis, was a reliable surrogate biomarker for pulmonary vascular resistance. We also used the same kind of empirical bayesian approach with another type of proteomic data, mass spectrometry data, in order to study differential analysis between different strains of Hepatitis C virus

[24]. These differential analyses were complemented by partial-least squares discriminant analysis. This last work was interesting not only for biology but also for Bilille platform to set up normalisation and statistical analysis pipelines for the PLBS P3M platform.

7.32 Axis 4: Testing Abnormality of a Sequence of Graphs: Application to Cybersecurity

Participants: Christophe Biernacki, Clarisse Boinay, Cristian Preda.

The increasing number of cyber attacks on industrial networks puts human life and economies at risk. Firms usually implement fixed rules rather than anomaly detection to prevent such attacks. However, anomaly detection methods would allow for a more flexible grasp of deviations from normal behaviour. For instance, anomaly detection in graphs modeling industrial networks can sense changes in the behaviour of machines. In this work, we seek to establish whether the number of messages sent from one or more machines to one or more machines is normal or not. To this end, we first model interactions between IP addresses with dynamical graphs. Then, we construct a test statistic based on the lihelihood of a graph computed thanks to generative models such as the stochastic block model and kernel estimators. Finally, we evaluate the power of the test in realistic and generic attack scenarios. This work has been presented to the main French conference in Statistics [38].

7.33 Axis 4:Bayesian spatiotemporal modelling for disease mapping: an application to preeclampsia and gestational diabetes in Florida, United States

Participants: Sophie Dabo-Niang.

Morbidities generally show patterns of concentration that vary by space and time. Disease mapping models are useful in estimating the spatiotemporal patterns of disease risks and are therefore pivotal for effective disease surveillance, resource allocation, and the development of prevention strategies. This study considers six spatiotemporal Bayesian hierarchical models based on two spatial conditional autoregressive priors. It could serve as a guideline on the development and application of Bayesian hierarchical models to assess the emerging risk trends, risk clustering, and spatial inequality trends, with estimation of covariables' effects on the interested disease risk. The method is applied to the Florida Birth Record data between 2006 and 2015 to study two cardiovascular risk factors: preeclampsia and gestational diabetes. High-risk clusters were detected in North Central Florida for preeclampsia and in Central Florida for gestational diabetes. While the adjusted disease trend was stable, spatial inequality peaked in 2011–2012 for both diseases. Exposure to PM2.5 at first or/and second trimester increased the risk of preeclampsia and gestational diabetes, but the magnitude is less severe compared to previous studies. In conclusion, this study underscores the significance of selecting appropriate disease mapping models in estimating the intricate spatiotemporal patterns of disease risk and suggests the importance of localized interventions to reduce health disparities. The result also identified an opportunity to study potential risk factors of preeclampsia, as the spike of risk in North Central Florida cannot be explained by current covariables.

This work is a result of a visit of Boubakari Ibrahimou (Florida International University, Miami, FL, USA) to Modal and university of Lille during two months.

It is a joint work with Ning Sun, Zoran Bursac, Ian Dryden, Roberto Lucchini, Boubakari Ibrahimou (Florida International University, Miami, FL, USA) [30].

7.34 Axis 4: Structural Changes in Temperature and Precipitation in MENA Countries

Participants: Sophie Dabo-Niang.

This paper evaluates the extent of climate variability in the Middle East and North Africa (MENA) region using time series structural change tests. The MENA region is highly susceptible to climate change, being one of the driest and most water-scarce regions in the world. The study aims to identify structural breaks in temperature and precipitation time series from 1901 to 2012. Specifically, a statistical analysis is performed based on a structural change model (Bai and Perron 1998, 2003a) for temperature and precipitation across 19 countries. The results indicate significant structural changes in temperature and precipitation patterns during the observation period, and suggest that climate variability has indeed begun to occur in all study area, with 1990 marking a turning point in terms of global warming. North African countries, Qatar, and the United Arab Emirates experienced a large number of breaks in temperature variables between 1901 and 2012, while other countries experienced fewer breaks. With regards to the seasonal aspect of precipitation, the individual rainfall Seasonality Index results demonstrate strong seasonal variability of rainfall from one year to another. Results show that rainfall in MENA countries is irregular throughout the year and that it ranges from seasonal to extremely seasonal throughout the study period. These findings have important implications for water resources management, agriculture, human health, and ecosystems in the region.

See for more details [12].

It is a joint work with Hassan Amouzay (University Mohamed V, Rabat), Raja Chakir (INRAE, Paris), Ahmed El Ghini (University Mohamed V, Rabat).

7.35 Axis 4: Spatial Relative Risk of Upper Aerodigestive Tract Cancers Incidence in French Northern Region

Participants: Sophie Dabo-Niang.

In this work, kernel spatial relative risk function estimation is of interest. We consider the case where covariates that may affect the spatial patterns of disease are contaminated by measurement errors. Finite sample properties were carried out in order to illustrate our methodology with real cancer data. We perform relative risk functions estimation on upper aerodigestive tract cancer (UADT) data to investigate locations of high and low incidence concentration in NPDC (Nord-Pas-de-Calais) French region.

For more details, see [17]. It is a joint work with Emad Darwich (University of Lille), Leila Hamdad, Hamid Haddadou (ESI, Algeria), Baba Thiam (University of Lille).

7.36 Axis 4: Functional, Multivariate Functional and Spatial PCA: Application to Covid-19 Data in the African Continent

Participants: Sophie Dabo-Niang.

Covid-19 pandemic has negatively impacted many areas, including the economy and health care facilities, and has left more than 5 million deaths worldwide. In this paper, we use functional data analysis methods to describe evolution of the number of cases and the number of deaths of Covid-19 in Africa.

We perform functional principal component analysis, Multivariate functional component analysis and spatial component analysis to characterize better the phenomena and spatial data to determine the impact of a region's neighborhood on number of cases. The obtained results allow us to have a better knowledge of the evolution of the pandemic in African continent.

It is a joint work with Idris Si-Ahmed (ESI, Algeria), Mazamaesso Azeyou (AIMS, Senegal), Leila Hamdad (ESI, Algeria). See for more details [67].

8 Bilateral contracts and grants with industry

8.1 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.

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

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

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.

Duration: 01/2021 - 12/2023 (3 years)

8.2 Bilateral grants with industry

Withings

Participants: Christophe Biernacki, Cristian Preda.

Withings is a French consumer electronics company which designs and innovates in connected devices, such as the first Wi-Fi scale on the market (introduced in 2009), an FDA-cleared blood pressure monitor, a smart sleep system, and a line of automatic activity tracking watches. It also provides B2B services for healthcare providers and researchers. A PhD program begun on September 2023 on the topic of analysis of multivariate, sparse longitudinal data, with mixed co-variates, from connected medical objects.

ADEO

Participants: Christophe Biernacki, Vincent Vandewalle.

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

Seckiot

Participants: Christophe Biernacki, Cristian Preda.

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

Decathlon

Participants: Cristian Preda.

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

Duration : 09/2022 - 08/2025 (3 years)

ASYGN

Participants: Sophie Dabo, Cristian Preda.

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

Duration: 01/2022 - 12/2024 (3 years)

HORIBA

Participants: Sophie Dabo, Cristian Preda.

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

Duration: 07/2021 - 12/2026 (6 years)

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

Since 2020, Benjamin Guedj is the founder and scientific director of the Inria London programme, an ambitious initiative to establish a joint research lab between Inria and University College London (UCL), framed within a broader bilateral Franco-British scientific cooperation.

9.2 International research visitors

9.2.1 Visits of international scientists

Participants: Sophie Dabo.

- Michelle Carey (University College Dublin) visited Sophie Dabo (July 2023, one week)
- Project title: Ireland and France Need Healthier Air for healthier Lungs: the Evidence
- PHC Ulysses 2023

Participants: Sophie Dabo.

- Sebastian Kuehnert (UC DAVIS, USA) visited Sophie Dabo (August-September 2023, one month)
- Project title: Functional Spatial Time series

9.2.2 Visits to international teams

Participants: Sophie Dabo.

Sophie Dabo visited

- University College Dublin (December 2023, one week)
- University of Malaya (Malaysia, July-August 2023, one month)
- University of Tokyo (Japan, August 2023, one week)
- North West University (South Africa, February 2023, one week)
- University of Vienna (Austria, April 2023, one week)
- University College London (March 2023, one week)

9.3 European initiatives

9.3.1 H2020 projects

H2020 FAIR

Participants: Guillemette Marot.

- Acronym: FAIR
- Project title: Flagellin aerosol therapy as an immunomodulatory adjunct to the antibiotic treatment of drug-resistant bacterial pneumonia
- Coordinator: JC Sirard (Inserm, CIIL)
- Duration: 5-6 years (2020-2025)

- Funding: 10 M euros
- 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)
- Contribution: 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 both as the scientific head of Bilille platform and as a researcher. At the beginning of the project, she has contributed to preliminary development of a tool to facilitate multi-omics data integration (visCorVar 6.1.4). In 2023, following reorganisation of the whole project by the coordinator, she has mostly contributed to longitudinal omics data analysis, by co-supervising an engineer with Pierre Pericard (Ulille, PLBS).

9.4 National initiatives

9.4.1 PEPR IA

Benjamin Guedj is a co-I of the project SHARP (PI: Rémi Gribonval, EP OCKHAM, CRI LYS) funded by the PEPR IA (2023-2027, overall funding 7M euros).

9.4.2 SIRIC EN-HOPE SMART4CBT

Participants: Sophie Dabo.

- Acronym: EN-HOPE SMART4CBT
- East North-Hematology Oncology Pediatric consortium offering a research program of Social sciences, Microenvironment and multiomics Analyses in RadioTherapy resistance For Children Brain Tumors
- Coordinator: ENTZ-WERLE Natacha (Inserm, University Hospital of Strasbourg)
- Duration: 5 years (2024-2028)
- Funding: 3M euros
- Partners: Inserm, CNRS, (France), Univ Lille, University Hospital of Nancy (CHRU Nancy), Oscar Lambret Centre in Lille, University Hospital of Lille (CHU Lille), ICANS, Institut du CANcer de Strasbourg Europe in Strasbourg ICL, Institut de Cancérologie de Lorraine in Nancy University of Strasbourg University of Lorraine
- Contribution: INCA

9.4.3 ANR

APRIORI

Participants: Benjamin Guedj.

- **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)

Synapark

Participants: Guillemette Marot.

- **Type:** ANR PRC
- **Acronym:** Synapark
- **Project title:** Evaluation of the role of parkin to alpha-synuclein-regulation in vitro, in vivo and in Parkinson's disease patient's blood samples
- **Coordinator:** Christine Alves da Costa (Inserm, IPMC)
- **Duration:** 42 months (2020–2024)
- **Funding:** 540k euros
- **Partners:** CNRS, Université Côte d'Azur, Univ. Lille, Inserm
- **Contribution:** Statistical analysis of transcriptomics data

CYTOMEMS

Participants: Sophie Dabo, Cristian Preda.

- **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)

Oesomics

Participants: Guillemette Marot.

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

TransEAsome

Participants: Guillemette Marot.

- **Type:** AMI Maladies rares
- **Acronym:** TransEAsome
- **Project title:** Long term outcome of esophageal atresia: transomics profiles in adolescence
- **Coordinator:** Frédéric Gottrand (Univ. Lille, CHU Lille, Infinite)
- **Duration:** 72 months (2022–2027)
- **Funding:** 1.4M euros
- **Partners:** CHU Lille, Univ. Lille, Inserm NO, Inserm ADR - GO, CRACMO, FIMATHO
- **Contribution:** Statistical analysis of multi-omics (mainly transcriptomics and proteomics) data

9.4.4 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:** 7 years (2016–2023)
- **Partners:** FHU Integra and Sanofi
- **Contribution:** 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 three 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.

- **Acronym:** PRECISE
- **Project title:** PREcision health in Complex Immune-mediated inflammatory diseaSEs
- **Coordinator:** David Launay (U. Lille, CHU Lille)
- **Duration:** 5 years (2021–2025)
- **Partners:** CHU Lille, CHU Amiens, CHU Rouen, CHU Caen, Université de Lille, Université de Picardie, Université de Rouen, Inserm
- **Contribution:** 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 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 platform in WP3 for the research of a molecular signature predictive of the treatment response (resistance and complication).

9.4.5 Inria 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), 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: 3 years (1/09/2021 - 31/12/2024)

9.4.6 Other national initiatives

Industrial Chair Smart digicat

Participants: 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.

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.

Duration: 6 years (1/07/2021 - 31/12/2026)

French Institute of Bioinformatics (IFB) and EquipEx+ MuDiS4LS

Participants: Guillemette Marot.

- **Coordinators:** IFB co-heads (changes in 2023)
- **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 the 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.

9.4.7 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 and the LEGO (machine learning for genomics) working groups.

9.5 Regional initiatives

Collaborations of the year linked to Bilille

Participants: Guillemette Marot.

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

- CIIL, Y. Rouillé

- LilNCog, D. Vieau
- SCALab, Y. Coello,
- OncoThai, K. Chen
- PLBS, J.-M. Saliou
- PRISM, T. Cardon

Collaborations of the year linked to ONCOLille

Participants: Sophie Dabo.

- SMMIL-E, C. Tarhan
- LIMMS, D. Collard
- Phycell, L. Lemonier,
- Canther, M. Cheock

10 Dissemination

Participants: Christophe Biernacki, Benjamin Guedj, Cristian Preda, Sophie Dabo, Guillemette Marot, Hemant Tiagy.

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

Christophe Biernacki and Hemant Tyagi organized on March 2023 a workshop dedicated to statistical learning on LARge scale GRaphs (LARGR) .

Guillemette Marot organized on July 2023 a workshop about numerical twins (both scientific and logistic organization).

Guillemette Marot organized on November 2023 two workshops related to scientific days of CNRS GDR BIM (Bioinformatique Moléculaire): - StatOmique (logistic and scientific organization), - LEGO (only logistic organization).

General chair, scientific chair Guillemette Marot was the scientific chair of the second session in the morning of LEGO workshop.

Member of the conference program committees Christophe Biernacki was member of an Inria/IIT DELHI workshop in New Delhi, related to the partnership between Inria and IIT Delhi. He gave also a talk on Digital Science for Disability for presenting and overview of the initiatives undertaken by Inria on this topic [33].

10.1.2 Journal

Member of the editorial boards Cristian Preda is an Associate Editor for Methodology and Computing in Applied Probability .

Benjamin Guedj is an Associate Editor for the journals JMLR, TMLR, Information and Inference, Data-Centric Engineering.

Reviewer - reviewing activities Christophe Biernacki acted as a reviewer for different journals (Statistics and Computing, Journal of Classification, Journal of Computational and Graphical Statistics...) and a conference (CAp 2023).

Guillemette Marot acted as a reviewer for ANR evaluation committee CE45 (Mathematics and Numerical sciences for biology and health)

Cristian Preda acted as a reviewer for Computational Statistics Journal.

Benjamin Guedj is a reviewer for JMLR, TMLR, Annals of Statistics, EJS, and most of the top-tier machine learning conferences (AISTATS, COLT, ICML, NeurIPS).

10.1.3 Invited talks

Christophe Biernacki was invited to give a plenary talk [36] and several all other talks [42], [35], [34], [57].

Hemant Tyagi gave a talk at the Inria/IIT Delhi workshop in New Delhi, and also an online talk at the City U. Hong Kong (Dept. of Mathematics).

10.1.4 Leadership within the scientific community

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

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

10.1.5 Scientific expertise

Cristian Preda gave a talk for Inria Academy program on generative models for articificial inteligence. See for more details Inria Academy program.

10.1.6 Research administration

Since January 2020, Christophe Biernacki acts as a deputy scientific director of Inria at the national level in charge of the domain “Applied mathematics, computation and simulation". Moreover, between October and December 2023, he was Director of the Inria research center at Lille (*intérim*).

Benjamin Guedj is the founder and scientific director of Inria London since 2020.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Christophe Biernacki gave four lessons on clustering for the “Ateliers tistiques de la SFdS” on June 2023 [65], [62], [63].

- Hemant Tyagi is teaching

  - Master: Statistics I, 24h, M1, Centrale Lille, France
  - Master: Statistics II, 24h, M1, Centrale Lille, France

- Sophie Dabo-Niang is teaching

  - Master: Spatial Statistics, 24h, M2, Université de Lille, France
  - Master: Advanced Statistics, 24h, M2, Université de Lille, France
  - Master: Multivariate Data Analyses, 24h, M2, Université de Lille, France
  - Licence: Probability, 24h, L2, Université de Lille, France
  - Licence: Multivariate Statistics, 24h, L3, Université de Lille, France

- Guillemette Marot is teaching

  - Licence: Biostatistics, 20h, L1, Université de Lille (Faculty of Medicine), France
  - Master: Biostatistics, 50h, M1, Université de Lille (Faculty of Medicine), France
  - Master: Supervised classification, 20h, M1, Polytech'Lille, France
  - Master: Biostatistics, 86h, M1, Université de Lille (Departments of Computer Science and Biology), France
  - Master: Artificial intelligence and health, M2, 3h, Université de Lille (Graduate school precision Health), France
  - Master: Statistical analysis of omic data, 15h, M2, Université de Lille (Department of Mathematics), France
  - Doctorat: Introduction to statistical analysis of omics data, 12h, Université de Lille (Faculty of Medicine), France

- Cristian Preda is teaching
  - Polytech'Lille engineer school: Linear Models, 48h.
  - Polytech'Lille engineer school: Advanced statistics, 48h.
  - Polytech'Lille engineer school: Biostatistics, 10h.
  - Polytech'Lille engineer school: Supervised clustering, 24h. France

- Benjamin Guedj is teaching
  - Probabilistic Modelling (M2, 30h), University College London, United Kingdom

10.2.2 Supervision

PhD in progress

- Axel Potier works on sale prediction for low turn-over products. Started in November 2020 under the supervision of Christophe Biernacki, Matthieu Marbac, Vincent Vandewalle.
- Clarisse Boinay works on anomaly detection and change point detection in contextual dynamic asynchronous graphs with applications in OT cybersecurity. Started in December 2021 under the supervision of Christophe Biernacki and Cristian Preda.
- Violaine Courrier works on the analysis of multivariate, sparse longitudinal data, with mixed covariates, from connected medical objects. Started in September 2023 under the supervision of Christophe Biernacki and Cristian Preda.
- Eglantine Karle works on dynamic ranking and translation synchronization on dynamic graphs. Started in November 2020 under the supervision of Hemant Tyagi.
- François Bassac works on functional data analysis for sport performance prediction. Started in September 2022 under the supervision of Cristian Preda.
- Reuben Adams works on PAC-Bayes theory. Started in September 2020 under the supervision of Benjamin Guedj.
- Clara Dubois works on functional data analysis with applications in Raman spectroscopy and chemical synthesis. Started in June 2023 under the supervision of Sophie Dabo.
- Antonin Schrab focuses on designing kernel-based hypothesis tests for two-sample comparisons. Started in September 2020 under the supervision of Benjamin Guedj.
- Maxime Haddouche works on statistical learning and PAC-Bayes theory. Started in September 2021 under the supervision of Benjamin Guedj.
- Etienne Kronert works on anomaly detection in time series. Started in September 2020 under the supervision of Alain Celisse.

10.2.3 Juries

- Christophe Biernacki acted as a reviewer for 5 PhD theses and for 1 HdR.
- Cristian Preda acted as a reviewer for 1 HDR.
- Benjamin Guedj acted as a reviewer for 3 PhD theses (Denmark, Germany, France).

11 Scientific production

11.1 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.

11.2 Publications of the year

International journals

[11] M.-S. Ahmed, M. N'diaye, M. K. Attouch and S. Dabo-Niang. 'k-nearest neighbors prediction and classification for spatial data'. In: *Journal of Spatial Econometrics* 4.1 (29th Nov. 2023), p. 12. DOI: 10.1007/s43071-023-00041-2. URL: https://hal.science/hal-04392790.

[12] H. Amouzay, R. Chakir, S. Dabo-Niang and A. El Ghini. 'Structural Changes in Temperature and Precipitation in MENA Countries'. In: *Earth Systems and Environment* 7.2 (6th May 2023), pp. 359–380. DOI: 10.1007/s41748-023-00344-2. URL: https://hal.science/hal-04392802.

[13] E. Araya, E. Karlé and H. Tyagi. 'Dynamic Ranking and Translation Synchronization'. In: *Information and Inference* 12.3 (27th Sept. 2023), pp. 2224–2266. DOI: 10.1093/imaiai/iaad029. URL: https://hal.science/hal-03876870.

[14] C. Biernacki, J. Jacques and C. Keribin. 'A Survey on Model-Based Co-Clustering: High Dimension and Estimation Challenges'. In: *Journal of Classification* (17th July 2023). URL: https://hal.science/hal-03769727.

[15] S. Bouka, S. Dabo-Niang and G. M. Nkiet. 'On estimation and prediction in spatial functional linear regression model'. In: *Lithuanian Mathematical Journal* 63.1 (10th Feb. 2023), pp. 13–30. DOI: 10.1007/s10986-023-09586-z. URL: https://hal.science/hal-04392795.

[16] G. Braun and H. Tyagi. 'Minimax Optimal Clustering of Bipartite Graphs with a Generalized Power Method'. In: *Information and Inference* 12.3 (27th Sept. 2023), pp. 1830–1866. DOI: 10.1093/imaiai/iaad006. URL: https://hal.science/hal-03876871.

[17] S. Dabo-Niang, E. Darwich, L. Hamdad and B. Thiam. 'Spatial Relative Risk of Upper Aerodigestive Tract Cancers Incidence in French Northern Region'. In: *SN Computer Science* 4.1 (Jan. 2023), p. 30. DOI: 10.1007/s42979-022-01426-0. URL: https://inria.hal.science/hal-03861906.

[18] C. Frévent, M.-S. Ahmed, S. Dabo-Niang and M. Genin. 'Investigating spatial scan statistics for multivariate functional data'. In: *Journal of the Royal Statistical Society: Series C Applied Statistics* 72.2 (12th May 2023), pp. 450–475. DOI: 10.1093/jrsssc/qlad017. URL: https://inria.hal.science/hal-03527471.

[19] Q. Grimonprez, S. Blanck, A. Celisse and G. Marot. 'MLGL: An R package implementing correlated variable selection by hierarchical clustering and group-Lasso'. In: *Journal of Statistical Software* 106.3 (23rd Mar. 2023). DOI: 10.18637/jss.v106.i03. URL: https://inria.hal.science/hal-01857242.

[20] M. Haddouche and B. Guedj. 'PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales'. In: *Transactions on Machine Learning Research Journal* (Apr. 2023). URL: https://inria.hal.science/hal-03815101.

[21] W. Heyse, V. Vandewalle, G. Marot, P. Amouyel, C. Bauters and F. Pinet. 'Identification of patient subtypes based on protein expression for prediction of heart failure after myocardial infarction'. In: *iScience* 26.3 (Mar. 2023), p. 106171. DOI: 10.1016/j.isci.2023.106171. URL: https://inserm.hal.science/inserm-04191363.

[22] E. Karlé and H. Tyagi. 'Dynamic Ranking with the BTL Model: A Nearest Neighbor based Rank Centrality Method'. In: *Journal of Machine Learning Research* 24.269 (23rd Sept. 2023), pp. 1–57. URL: https://inria.hal.science/hal-03519271.

[23] T. H. Khoo, D. Pathmanathan and S. Dabo-Niang. 'Spatial Autocorrelation of Global Stock Exchanges Using Functional Areal Spatial Principal Component Analysis'. In: *Mathematics* 11.3 (28th Jan. 2023), p. 674. DOI: 10.3390/math11030674. URL: https://hal.science/hal-04392798.

[24] E. Martin de Fourchambault, N. Callens, J.-M. Saliou, M. Fourcot, O. Delos, N. Barois, Q. Thorel, S. Ramirez, J. Bukh, L. Cocquerel, J. Bertrand-Michel, G. Marot, Y. Sebti, J. Dubuisson and Y. Rouillé. 'Hepatitis C virus alters the morphology and function of peroxisomes'. In: *Frontiers in Microbiology* 14 (21st Sept. 2023), p. 1254728. DOI: 10.3389/fmicb.2023.1254728. URL: https://hal.science/hal-04245832.

[25] I.-A. Moindjié, S. Dabo-Niang and C. Preda. 'Classification of multivariate functional data on different domains with Partial Least Squares approaches'. In: *Statistics and Computing* (19th Oct. 2023), p. 5. DOI: 10.1007/s11222-023-10324-1. URL: https://hal.science/hal-03908634.

[26] V. Raverdy, E. Chatelain, G. Lasailly, R. Caiazzo, J. Vandel, H. Verkindt, C. Marciniak, B. Legendre, P. Bauvin, N. Oukhouya-Daoud, G. Baud, M. Chetboun, M.-C. Vantyghem, V. Gnemmi, E. Leteurtre, B. Staels, P. Lefebvre, P. Mathurin, G. Marot and F. Pattou. 'Combining diabetes, sex, and menopause as meaningful clinical features associated with NASH and liver fibrosis in individuals with class II and III obesity: A retrospective cohort study.' In: *Obesity*. Obesity 31 (21st Nov. 2023), pp. 3066–3076. DOI: 10.1002/oby.23904. URL: https://hal.univ-lille.fr/hal-04402748.

[27] S. Sanges, L. Rice, L. Tu, E. Valenzi, J.-L. Cracowski, D. Montani, J. Mantero, C. Ternynck, G. Marot, A. Bujor, E. Hachulla, D. Launay, M. Humbert, C. Guignabert and R. Lafyatis. 'Biomarkers of haemodynamic severity of systemic sclerosis-associated pulmonary arterial hypertension by serum proteome analysis'. In: *Annals of the Rheumatic Diseases* (10th Feb. 2023), ard-2022–223237. DOI: 10.1136/ard-2022-223237. URL: https://inria.hal.science/hal-03927525.

[28] H. Sarter, G. Savoye, G. Marot, D. Ley, D. Turck, J.-P. Hugot, F. Vasseur, A. Duhamel, P. Wils, F. Princen et al. 'A Novel 8-Predictors Signature to Predict Complicated Disease Course in Pediatric-onset Crohn's Disease: A Population-based Study'. In: *Inflammatory Bowel Diseases* (2nd June 2023). DOI: 10.1093/ibd/izad090. URL: https://u-picardie.hal.science/hal-04124611.

[29] A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj and A. Gretton. 'MMD Aggregated Two-Sample Test'. In: *Journal of Machine Learning Research* 24.194 (June 2023), pp. 1–81. URL: https://inria.hal.science/hal-03408976.

[30] N. Sun, Z. Bursac, I. Dryden, R. Lucchini, S. Dabo-Niang and B. Ibrahimou. 'Bayesian spatiotemporal modelling for disease mapping: an application to preeclampsia and gestational diabetes in Florida, United States'. In: *Environmental Science and Pollution Research* 30.50 (28th Sept. 2023), pp. 109283–109298. DOI: 10.1007/s11356-023-29953-0. URL: https://hal.science/hal-04392782.

[31] R. T. Tchouya, S. Nasini and S. Dabo-Niang. 'An estimation approach for the influential–imitator diffusion'. In: *Computers and Operations Research* 159 (Nov. 2023), p. 106315. DOI: 10.1016/j.cor.2023.106315. URL: https://hal.science/hal-04392806.

International peer-reviewed conferences

[32] C. Biernacki. 'Levels Merging in the Latent Class Model'. In: Statistical Learning Sustainability and Impact Evaluation. Ancona (IT), Italy, 21st June 2023. URL: https://inria.hal.science/hal-04370900.

[33] C. Biernacki. 'Digital Science for Disability Overview of the initiatives undertaken by Inria'. In: Inria/IIT DELHI workshop. New delhi, India, 25th Oct. 2023. URL: https://inria.hal.science/hal-04370905.

[34] C. Biernacki. 'Levels Merging in the Latent Class Model'. In: CFE-CMStatistics 2023. Berlin (Germany), Germany, 16th Dec. 2023. URL: https://inria.hal.science/hal-04370783.

[35] C. Biernacki, C. Boyer, G. Celeux, J. Josse, F. Laporte, M. Marbac, A. Sportisse and V. Vandewalle. 'Impact of missing data on mixtures and clustering with illustrations in Biology and Medicine'. In: SPSR 2023 - The 24th annual Conference of the Romanian Society of Probability and Statistics. Bucarest, Romania, 21st Apr. 2023. URL: https://inria.hal.science/hal-04370870.

[36] C. Biernacki, V. Vandewalle and M. Marbac. 'Clustering: from modeling to visualizing Mapping clusters as spherical Gaussians'. In: SFC 2023 - Rencontres de la Société Francophone de Classification. Rencontres de la Société Francophone de Classification. Strasbourg, France, 6th July 2023. URL: https://inria.hal.science/hal-04370886.

[37] F. Biggs and B. Guedj. 'Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty'. In: AISTATS 2023 - 26th International Conference on Artificial Intelligence and Statistics. Valencia, Spain, 20th Oct. 2022. URL: https://inria.hal.science/hal-03953307.

[38] C. Boinay, C. Biernacki and C. Preda. 'Graphs in OT Testing Graph Abnormality Application to a Real OT Data Set Ongoing & Future Works References'. In: 54e Journées de Statistique. Bruxelles, Belgium, 3rd July 2023. URL: https://inria.hal.science/hal-04370878.

[39] P. Viallard, M. Haddouche, U. Şimşekli and B. Guedj. 'Learning via Wasserstein-Based High Probability Generalisation Bounds'. In: NeurIPS 2023 - Thirty-seventh Conference on Neural Information Processing Systems. New Orleans, United States, 7th June 2023. DOI: 10.48550/arXiv.2306.04375. URL: https://hal.science/hal-04121624.

[40] P. Viallard, M. Haddouche, U. Şimşekli and B. Guedj. 'Learning via Wasserstein-Based High Probability Generalisation Bounds'. In: NeurIPS 2023 Workshop on Optimal Transport and Machine Learning (OTML'23). New Orleans, United States, 16th Dec. 2023. URL: https://hal.science/hal-04273718.

National peer-reviewed Conferences

[41] V. Courrier, C. Biernacki, C. Preda and B. Vittrant. 'Comparative study of clustering models for multivariate time series from connected medical devices'. In: EGC 2024 - 24ème Conférence Francophone sur l'Extraction et Gestion des Connaissances. Dijon, France, 23rd Jan. 2024. URL: https://riip.hal.science/pasteur-04364645.

Conferences without proceedings

[42] C. Biernacki, M. Marbac and V. Vandewalle. 'Clustering: from modeling to visualizing Mapping clusters as spherical Gaussians'. In: Working Group on Model-Based Clustering. Pittsburg, Etats-Unis, United States, 17th July 2023. URL: https://inria.hal.science/hal-04370893.

[43] C. Preda. 'Learning with categorical functional data'. In: The Tenth Congress of Romanian Mathematicians, 2023. Pitesti, Romania, 30th June 2023. URL: https://hal.science/hal-04393613.

[44] C. Preda and Q. Grimonprez. 'Linear approximation for multivariate categorical functional data analysis'. In: THE 24th CONFERENCE of the ROMANIAN SOCIETY of PROBABILITY and STATISTICS. Bucharest, Romania, 21st Apr. 2023. URL: https://hal.science/hal-04393621.

Doctoral dissertations and habilitation theses

[45] I.-A. Moindjié. 'Linear models for multivariate functional data'. Université de Lille, 18th Dec. 2023. URL: https://hal.science/tel-04376932.

Reports & preprints

[46] F. Biggs, A. Schrab and A. Gretton. *MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting*. 15th June 2023. URL: https://hal.science/hal-04156329.

[47] E. Clerico and B. Guedj. *A note on regularised NTK dynamics with an application to PAC-Bayesian training*. 20th Dec. 2023. URL: https://inria.hal.science/hal-04358260.

[48] M. Haddouche and B. Guedj. *Wasserstein PAC-Bayes Learning: Exploiting Optimisation Guarantees to Explain Generalisation*. 14th Apr. 2023. URL: https://inria.hal.science/hal-04080080.

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

[50] F. Hellström, G. Durisi, B. Guedj and M. Raginsky. *Generalization Bounds: Perspectives from Information Theory and PAC-Bayes*. 8th Sept. 2023. URL: https://inria.hal.science/hal-04205539.

[51] F. Hellström and B. Guedj. *Comparing Comparators in Generalization Bounds*. 16th Oct. 2023. URL: https://inria.hal.science/hal-04259101.

[52] P. Jobic, M. Haddouche and B. Guedj. *Federated Learning with Nonvacuous Generalisation Bounds*. 17th Oct. 2023. URL: https://inria.hal.science/hal-04260486.

[53] I. Kim and A. Schrab. *Differentially Private Permutation Tests: Applications to Kernel Methods*. 31st Oct. 2023. URL: https://hal.science/hal-04276141.

[54] E. Krönert, A. Célisse and D. Hattab. *FDR control for Online Anomaly Detection*. 1st Dec. 2023. URL: https://hal.science/hal-04321622.

[55] I.-A. Moindjié, C. Preda and S. Dabo-Niang. *Fusion regression methods with repeated functional data*. 28th Nov. 2023. URL: https://hal.science/hal-04176783.

[56] A. Schrab, W. Jitkrittum, Z. Szabó, D. Sejdinovic and A. Gretton. *Discussion of 'Multiscale Fisher's Independence Test for Multivariate Dependence'*. 8th Nov. 2023. URL: https://hal.science/hal-03709218.

[57] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, C. Biernacki and J. Josse. *Accompanying note : Model-based Clustering with Missing Not At Random Data.* 21st Dec. 2023. URL: https://hal.science/hal-04358192.

[58] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, J. Josse and C. Biernacki. *Model-based Clustering with Missing Not At Random Data.* 21st Dec. 2023. URL: https://hal.science/hal-03494674.

[59] H. Tyagi and D. Efimov. *Learning linear dynamical systems under convex constraints.* 2nd Aug. 2023. URL: https://hal.science/hal-04394297.

[60] E. A. Valdivia and H. Tyagi. *Graph Matching via convex relaxation to the simplex.* 1st Nov. 2023. URL: https://hal.science/hal-04394319.

Other scientific publications

[61] J. Dubois-chevalier, C. Gheeraert, A. Berthier, C. Boulet, V. Dubois, L. Guille, M. Fourcot, G. Marot, K. Gauthier, L. Dubuquoy, B. Staels, P. Lefebvre and J. Eeckhoute. 'An extended transcription factor regulatory network controls hepatocyte identity'. In: *EMBO Reports* (10th July 2023). DOI: 10.15252/embr.202357020. URL: https://hal.science/hal-04169341.

11.3 Other

Educational activities

[62] C. Biernacki. 'Clustering : une vision unifiée pour une utilisation éclairée - Partie 2 Evaluation d'une méthode de clustering'. Doctoral. France, 15th June 2023. URL: https://inria.hal.science/hal-04370753.

[63] C. Biernacki. 'Clustering : une vision unifiée pour une utilisation éclairée - Partie 3 : Formalisation par modèles de mélange'. Doctoral. France, 16th June 2023. URL: https://inria.hal.science/hal-04370761.

[64] C. Biernacki. 'Clustering : une vision unifiée pour une utilisation éclairée - Partie 4 : Traitement de la grande dimension & co-clustering'. Doctoral. France, 16th June 2023. URL: https://inria.hal.science/hal-04370767.

[65] C. Biernacki. 'Clustering : une vision unifiée pour une utilisation éclairée - Partie 1 : Méthodes exploratoires en clustering'. Doctoral. France, 15th June 2023. URL: https://inria.hal.science/hal-04370746.

11.4 Cited publications

[66] C. J. Agonkoui, F. D. Moussa and S.-D. Niang. 'Multivariate functional principal component analysis for endogenously stratified data'. In: *Afrika Statistika* 17.4 (Oct. 2022), pp. 3321–337. DOI: 10.16929/as/2022.3321.308. URL: https://hal.science/hal-04392805.

[67] I. Si-Ahmed, M. Azeyou, L. Hamdad and S. Dabo-Niang. 'Functional, Multivariate Functional and Spatial PCA: Application to Covid-19 Data in the African Continent'. In: *12th International Conference on Information Systems and Advanced Technologies ICISAT 2022.* Vol. 624. Lecture Notes in Networks and Systems. Virtual conference, France: Springer International Publishing, Aug. 2022, pp. 318–328. DOI: 10.1007/978-3-031-25344-7_28. URL: https://hal.science/hal-04392801.