RESEARCH CENTRE

Inria Centre at the University of Lille

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2024

ACTIVITY REPORT

Team

MODAL

MOdel for Data Analysis and Learning

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions)

IN COLLABORATION WITH: Laboratoire Paul Painlevé (LPP)

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Inria

Contents

Team MODAL

Creation of the Team: 2024 January 01

Keywords

Computer sciences and digital sciences

A3.1.4. – Uncertain data

A3.1.10. – Heterogeneous data

A3.2.3. – Inference

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.5. – Bayesian methods

A3.4.7. – Kernel methods

A5.2. – Data visualization

A5.9.2. – Estimation, modeling

A6.2.3. – Probabilistic methods

A6.2.4. – Statistical methods

A6.3.3. – Data processing

A9.2. – Machine learning

Other research topics and application domains

B2.2.3. – Cancer

B9.5.6. – Data science

B9.6.3. – Economy, Finance

B9.6.5. – Sociology

1 Team members, visitors, external collaborators

Research Scientists

- Christophe Biernacki [INRIA, Senior Researcher, from Sep 2024]
- Christophe Biernacki [INRIA, Professor Detachement, until Aug 2024]
- Benjamin Guedj [INRIA, Senior Researcher, from Oct 2024]
- Benjamin Guedj [INRIA, Researcher, until Sep 2024]
- Hemant Tyagi [INRIA, Researcher]

Faculty Members

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

Post-Doctoral Fellows

- Christelle Agonkoui [INRIA, Post-Doctoral Fellow, from Sep 2024]
- Gaurav Dhar [INRIA, Post-Doctoral Fellow, from Nov 2024]
- Komlan Midodzi Noukpoape [UNIV LILLE, Post-Doctoral Fellow, from May 2024]

PhD Students

- François Bassac [DECATHLON, CIFRE]
- Clarisse Boinay [Seckiot, CIFRE, until Nov 2024]
- Hugo Cannafarina [INRIA, from Nov 2024]
- Violaine Courrier [WITHINGS, CIFRE]
- Clara Dubois [LABO TIMC, CIFRE]
- Maxime Haddouche [UNIV LILLE, until Sep 2024]
- Antoine Picard [UCL, CIFRE]
- Axel Potier [ADEO, CIFRE, until Mar 2024]

Technical Staff

- Louise Chen [INRIA, Engineer, until Oct 2024]
- Axel Potier [INRIA, Engineer, from Apr 2024 until Jun 2024]

Interns and Apprentices

- Hugo Cannafarina [ENSAI, Intern, from Apr 2024 until Sep 2024]
- Nicolas Desmons [INRIA, Intern, from Jul 2024 until Sep 2024]

Administrative Assistant

- Anne Rejl [INRIA]

External Collaborator

- Alain Celisse [UNIV PARIS I]

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 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 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 expectation-maximization 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 Highlights of the year

6.1 Awards

- Christophe Biernacki was elected as Head of the SFdS (Société Française de Statistique) since July 2024, which is the French society specialized in statistics, whose mission is to promote the use of statistics and its understanding and to foster its methodological developments.
- Sophie Dabo got a CNRS chair (2024-2027) within a joint research program with Africa.
- Guilemette Marot is co-author of the paper "Data-driven cluster analysis identifies distinct types of metabolic dysfunction-associated steatotic liver disease", published in Nature Medicine.
- Sophie Dabo was elected on June 2024 vice-president of CIMPA (International Center of Pure and Applied Mathematics), whose mission is to promote the developments of Mathematics in the developping world. The nomination will be made official during the steering committee of January 2025.

7 New software, platforms, open data

7.1 New software

7.1.1 MixtComp.V4

Keyword: Clustering

Functional Description: MixtComp (Mixture Computation) is a model-based clustering package for mixed data from 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

7.1.2 cfda

Name: Categorical functional data analysis

Keyword: Functional data

Functional Description: The R package cfda performs:

- descriptive statistics for categorical functional data

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

- approximation for multivariate categorical functional data analysis.

Release Contributions: - approximation for multivariate categorical functional data analysis.

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

Contact: Cristian Preda

Participants: Cristian Preda, Quentin Grimonprez, Vincent Vandewalle

Partner: Université de Lille

7.1.3 ClusPred

Name: Simultaneous Semi-Parametric Estimation of Clustering and Regression

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

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

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

Contact: Matthieu Marbac-Lourdelle

7.1.4 visCorVar

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

Keywords: Data integration, Visualization

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

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

Contact: Guillemette Marot

Participants: Maxime Brunin, Guillemette Marot, Pierre Pericard

Partner: Université de Lille

7.1.5 metaRNASeq

Name: RNA-Seq data meta-analysis

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

7.1.6 HDSpatialScan

Name: Multivariate and Functional Spatial Scan Statistics

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

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

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

Contact: Sophie Dabo

7.1.7 MLGL

Name: Multi-Layer Group Lasso

Keywords: Variable selection, Statistical learning

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

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

Contact: Guillemette Marot

8 New results

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

Participants: Christophe Biernacki.

Model-based unsupervised learning, as any learning task, stalls as soon as missing data occurs. This is even more true when the missing data are informative, or said missing not at random (MNAR). We propose model-based clustering algorithms designed to handle very general types of missing data, including MNAR data. To do so, we introduce a mixture model for different types of data (continuous, count, categorical and mixed) to jointly model the data distribution and the MNAR mechanism, remaining vigilant to the relative degrees of freedom of each. Several MNAR models are discussed, for which the cause of the missingness can depend on both the values of the missing variable themselves and on the class membership. However, we focus on a specific MNAR model, called MNARz, for which the missingness only depends on the class membership. We first underline its ease of estimation, by showing that the statistical inference can be carried out on the data matrix concatenated with the missing mask considering finally a standard MAR mechanism. Consequently, we propose to perform clustering using the Expectation Maximization algorithm, specially developed for this simplified reinterpretation. Finally, we assess the numerical performances of the proposed methods on synthetic data and on the real medical registry TraumaBase as well.

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-Lourdelle from ENSAI.

See for more details [26].

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

Participants: Christophe Biernacki.

Standard model-based clustering is known to be very efficient for low-dimensional data sets, but it fails for properly addressing high dimension (HD) ones, where it suffers from both statistical and computational drawbacks. In order to counterbalance this curse of dimensionality, some proposals have been made to take into account redundancy and features utility, but related models are not suitable for too many variables. We advocate that co-clustering, an unsupervised mixture model learning method to define simultaneously groups of rows (individuals) and groups of columns (variables) on a data matrix, 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. However, the statistical counterpart of this important variance reduction brings naturally some important model bias.

A presentation at an international conference [29] 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.

8.3 Axis 1: An EM Stopping Rule for Avoiding Degeneracy in Gaussian-based Clustering with Missing Data

Participants: Christophe Biernacki.

Missing data frequency increases with the growing size of multivariate modern datasets. In Gaussian model-based clustering, the EM algorithm easily takes into account such data but the degeneracy problem is dramatically aggravated during the EM runs: parameter degeneracy is quite slow and also more frequent than with complete data. Consequently, parameter degenerated solutions may be confused with valuable parameter solutions and, in addition, computing time may be wasted through wrong runs. In this work, a simple and low informational condition on the latent partition allows to propose a very simple partition-based stopping rule of EM which shows good behavior on numerical experiments.

This work has been presented to an international conference [30], to a national conference [28] and also to the "séminaire parisien de statistique" [44].

It is a joint work with Vincent Vandewalle from University Côte d'Azur.

8.4 Axis 4: Estimating Substitution for Optimised Replenishment with Slow Movers Products

Participants: Christophe Biernacki, Axel Potier.

In Retail, inventory optimisation is a common problem targeting a trade-off between the risk of stock-out and the risk of overstocking, in order to reach an optimal global profit. This inventory optimisation task is however very challenging in case of products that are sold in low quantity (so-called slow movers in Retail). Indeed, estimating properly the related future sales, especially after discretisation, usually suffers from high relative standard deviation, from which the optimal replenishment solution inherits to the point of being usefulness in practice. Nevertheless, slow movers in many companies, as ADEO (French holding company selling consumer goods for decoration), are sufficiently numerous to represent a significant portion of the sales and of the stock. Consequently, concerned items are difficult to be optimally replenished, and even a small improvement of the replenishment process may have a significant positive effect on the whole global profit. For bridging the gap, we reformulate the slow movers optimal replenishment problem as a substitution probability estimation problem between items. When a product is out of stock, a client may alternatively choose another item (a so-called substitute product), avoiding to definitively loose the intended initial sale. As a consequence, instead of choosing the optimal quantity to replenish separately for each item (classical approach), we leverage that additional information that a group of items products can be substituted for each other allows to more efficiently estimate the optimal replenished quantity of the whole group of items. Obviously, such a substitution process occurs only with a certain probability, (1) that we have to estimate and (2) then we have to properly use through the optimal replenishment calculation. Notice fundamentally that the discrete nature of the stock quantity is expected to provide a better relative improvement in terms of profit in the case of slow movers compared to the case of fast movers (the contrary of slow movers), which justifies the special interest of our approach for slow movers.

This thesis has been defended this year [36] and a associated preprint is under preparation for submission to an international journal of statistics.

It is a joint work with Vincent Vandewalle from University Côte d'Azur, Matthieu Marbac-Lourdelle from ENSAI and Julien Favre from the ADEO company.

8.5 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 a specialized seminar in Cybersecurity at Campus Cyber – La Défense [43].

8.6 Axis 4: Comparative study of clustering models of multivariate time series from connected medical objects

Participants: Christophe Biernacki, Violaine Courrier, Cristian Preda.

In healthcare, patient data are often collected as multivariate time series, providing a comprehensive view of a patient's health status over time. These data are typically sparse and episodic. However, connected medical devices can increase the frequency of data. The goal is to create patient profiles from these time series in an unsupervised manner. In the absence of labels, a predictive model can be used to predict future values while forming a latent cluster space, evaluated based on predictive performance. Using real-world data from Withings, we compare the static clustering approaches MAGMACLUST, which creates a cluster across the entire time series, and dynamic clustering approaches DGM2, which allows an individual's membership in a group to change over time.

This work has been presented to a national conference [33].

It is a joint work with Benjamin Vittrant from the Withings company.

8.7 Axis 2: Anomaly detection in time series using breakpoint detection and multiple testing

Participants: Etienne Krönert, Alain Celisse.

The purpose of this work is to develop new methods in the field of anomaly detection. An anomaly detector should identify data points that have been generated by a process different from the reference process. In general, anomaly detectors are trained on past data and only consider the reference distribution given in the training set. Alternatively, the detector is updated in real time from a fixed length sliding window. Neither approach takes into account the true dynamics of the time series, leading to false positives when the reference distribution changes.The proposed approach consists in detecting these distribution changes upstream, using breakpoint detectors. Anomalies are then retrieved within the identified homogeneous segments.The advantage of this approach for a company like Worldline is that it can quickly detect incidents in its system, while reducing the number of false alarms.In addition, care is taken to theoretically control the number of false positives through the False Detection Rate (FDR).The first step is an attempt to develop a new estimator of the p-value. After comparison with existing estimators, this new estimator turns out to be too complex, with no real gain. The empirical p-value estimator is then used.Then, the problem of anomaly detection on iid series is studied theoretically. A procedure is developed to control the FDR of a series of infinite length by controlling a variant of the FDR on subseries of fixed length. Empirical experiments show under which conditions this control is achieved in practice. Finally, our new anomaly detector based on breakpoints is introduced for piecewise iid series. It is shown that the FDR control procedure is effective in this new context. However, the use of a breakpoint detector leads to two difficulties: small segment length and breakpoint detection delays. To overcome these

difficulties, a confidence score is introduced. The detector is empirically evaluated to show the relevance and limitations of our approach. See for more details [35].

8.8 Axis 2: Breakpoint based online anomaly detection

Participants: Etienne Krönert, Alain Celisse.

This work proposes a new online anomaly detector for time series. Classically, anomaly detectors do not adapt well in real time to changes in the reference distribution. The novelty of our approach is to use breakpoint detection to adapt online to the new reference behavior of the time series. The statistical performance of the detector is theoretically ensured by a control on the FDR. The anomaly detector is empirically evaluated in depth to assess its capabilities and limitations. See for more details [39].

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

Participants: Ernesto Araya, Guillaume Braun, Hemant Tyagi.

In the graph matching problem we observe two graphs G, H and the goal is to find an assignment (or matching) between their vertices such that some measure of edge agreement is maximized. We assume in this work that the observed pair G, H has been drawn from the Correlated Gaussian Wigner (CGW) model – a popular model for correlated weighted graphs – where the entries of the adjacency matrices of G and H are independent Gaussians and each edge of G is correlated with one edge of H (determined by the unknown matching) with the edge correlation described by a parameter $\sigma \in [0,1)$. In this paper, we analyse the performance of the projected power method (PPM) as a seeded graph matching algorithm where we are given an initial partially correct matching (called the seed) as side information. We prove that if the seed is close enough to the ground-truth matching, then with high probability, PPM iteratively improves the seed and recovers the ground-truth matching (either partially or exactly) in $\mathcal{O}(\log n)$ iterations. Our results prove that PPM works even in regimes of constant σ, thus extending the analysis in (Mao et al.,2021) for the sparse Correlated Erdös-Renyi (CER) model to the (dense) Wigner model. As a byproduct of our analysis, we see that the PPM framework generalizes some of the state-of-art algorithms for seeded graph matching. We support and complement our theoretical findings with numerical experiments on synthetic data.

This work has been published in the Journal of Machine Learning Research (JMLR) ([12]).

8.10 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 appeared in the journal Foundations of Data Science [12].

8.11 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 [42] and is joint work with Denis Efimov (Inria Lille, Valse team).

8.12 Axis 2: Joint Learning of Linear Dynamical Systems under Smoothness Constraints

Participants: Hemant Tyagi.

We consider the problem of joint learning of multiple linear dynamical systems. This has received significant attention recently under different types of assumptions on the model parameters. The setting we consider involves a collection of m linear systems each of which resides on a node of a given undirected graph $G = ([m], \mathcal{E})$. We assume that the system matrices are marginally stable, and satisfy a smoothness constraint w.r.t G – the smoothness measured in a manner akin to the quadratic variation of a signal on a graph. Given access to the states of the nodes over T time points, we then propose two estimators for joint estimation of the system matrices, along with non-asymptotic error bounds on the mean-squared error (MSE). In particular, we show conditions under which the MSE converges to zero as m increases, typically polynomially fast w.r.t m. The results hold under mild (i.e., $T \sim \log m$), or sometimes, even no assumption on T (i.e. $T \geq 2$).

This work is currently under review in a journal [41].

8.13 Axis 1&2: Dynamic angular synchronization under smoothness constraints

Participants: Ernesto Araya, Mihai Cucuringu, Hemant Tyagi.

Given an undirected measurement graph $\mathcal{H} = ([\backslash], \mathcal{E})$, the classical angular synchronization problem consists of recovering unknown angles $\theta_{1^*}, \ldots, \theta_{n^*}$ from a collection of noisy pairwise measurements of the form $(\theta_i^* - \theta_j^*)\ mod\ 2\pi$, for all $i, j \in \mathcal{E}$. This problem arises in a variety of applications, including computer vision, time synchronization of distributed networks, and ranking from pairwise comparisons.

In this paper, we consider a dynamic version of this problem where the angles, and also the measurement graphs evolve over T time points. Assuming a smoothness condition on the evolution of the latent angles, we derive three algorithms for joint estimation of the angles over all time points. Moreover, for one of the algorithms, we establish non-asymptotic recovery guarantees for the mean-squared error (MSE) under different statistical models. In particular, we show that the MSE converges to zero as T increases under milder conditions than in the static setting. This includes the setting where the measurement graphs are highly sparse and disconnected, and also when the measurement noise is large and can potentially increase with T. We complement our theoretical results with experiments on synthetic data.

This work is currently under review in a journal [37].

8.14 Axis 2: A shared-frailty spatial scan statistic model for time-to-event data

Participants: Sophie Dabo-Niang.

This paper presents a scan statistic based on a Cox model with shared frailty that takes into account the spatial correlation between spatial units. In simulation studies, we have shown that (i) classical models of spatial scan statistics for time-to-event data fail to maintain the type I error in the presence of intra-spatial unit correlation, and (ii) our model performs well in the presence of both intra-spatial unit correlation and inter-spatial unit correlation. Our method has been applied to epidemiological data and to the detection of spatial clusters of mortality in patients with end-stage renal disease in northern France.

It is a joint work with with Camille Frévent, Mohamed Salem Ahmed and Michael Genin (University of Lille, CERIM) [17].

8.15 Axis 2: A Markov-switching spatio-temporal ARCH model

Participants: Sophie Dabo-Niang.

This paper proposes a Markov switching framework of the spatio-temporal log-ARCH model. We investigate the estimation procedure and the smooth inferences of the regimes. The Monte-Carlo simulation studies show that the maximum likelihood estimation method for our proposed model has good finite sample properties. The proposed model was applied to 28 stock indices data that were presumably affected by the 2015-2016 Chinese stock market crash. The results showed that our model is a better fit compared to that of the one-regime counterpart. Furthermore, the smoothed inference of the data indicated the approximate periods where structural breaks occurred. This model can capture structural breaks that simultaneously occur in nearby locations.

It is a joint work with with Tzung Hsuen Khoo, Dharini Pathmanathan (Malaya University, Malaysia), Philipp Otto (University of Glasgow) [20].

8.16 Axis 2: Automatic geomorphological mapping using ground truth data with coverage sampling and random forest algorithms

Participants: Sophie Dabo-Niang.

This work provides statistical learning approaches based framework to build automatically geomorphological maps. We used bathymetric data to build Digital Bathymetric Model (DBM) and compute terrain attributes characteristic of seafloor geomorphology. Then, we used clustering based algorithms to select automatically ground truth locations from a reference geomorphological map manually made. Finally a supervised classification model random forest based was used to build predictive models for

seafloor geomorphology typologies. Subsequently we studied the effect of DBM resolution, sample size and sampling method of the ground truth locations, in the quality of map production via a series of simulations. Results showed that the proposed framework allowed to build efficiently relevant seafloor geomorphological maps. The best compromise between the sampling effort and the quality of the resulting maps was obtained with 100 m DBM resolution, 200 data points sample size and using a complexity-dependent sampling method and led to a map matching at 90% the reference one.

It is a joint work with with Paul Aimé Latsouck Faye, Elodie Brunel, Thomas Claverie, Solym Mawaki Manou-Abi (University of Montpellier 2) [16].

8.17 Axis 2: FDR control for Online Anomaly Detection

Participants: Etienne Krönert, 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 [38].

8.18 Axis 3: Uncovering Data Across Continua : An Introduction to Functional Data Analysis

Participants: Sophie Dabo-Niang.

This article introduces FDA, merging statistics with real-world complexity, ideal for those with mathematical skills but no FDA background.

It is a joint work with Camille Frévent (University of Lille, METRICS). It bas been published in the AMS (American Mathematical Society) Notices ([14]).

8.19 Axis 3: Principal Component Analysis of Multivariate Spatial Functional Data

Participants: Sophie Dabo-Niang.

This paper is devoted to the study of dimension reduction techniques for multivariate spatially indexed functional data and defined on different domains. We present a method called Spatial Multivariate Functional Principal Component Analysis (SMFPCA), which performs principal component analysis for multivariate spatial functional data. In contrast to Multivariate Karhunen-Loève approach for independent data, SMFPCA is notably adept at effectively capturing spatial dependencies among multiple functions. SMFPCA applies spectral functional component analysis to multivariate functional spatial data, focusing on data points arranged on a regular grid. The methodological framework and algorithm of SMFPCA have been developed to tackle the challenges arising from the lack of appropriate methods for managing this type of data. The performance of the proposed method has been verified through finite sample properties using simulated datasets and sea-surface temperature dataset. Additionally, we

conducted comparative studies of SMFPCA against some existing methods providing valuable insights into the properties of multivariate spatial functional data within a finite sample.

It is a joint work with IdrisSi-ahmeda, Leila Hamdad (ESI, Algeria) Christelle Judith Agonkoui (IMSP, Benin) and Yoba Kande (IRD, Senegal and University of Lille). For more details, see [11].

8.20 Axis 3: Forecasting mortality rates with functional signatures

Participants: Sophie Dabo-Niang.

This study introduces an innovative methodology for mortality forecasting, which integrates signature-based methods within the functional data framework of the Hyndman-Ullah (HU) model. This new approach, termed the Hyndman-Ullah with truncated signatures (HUts) model, aims to enhance the accuracy and robustness of mortality predictions. By utilizing signature regression, the HUts model is able to capture complex, nonlinear dependencies in mortality data which enhances forecasting accuracy across various demographic conditions. The model is applied to mortality data from 12 countries, comparing its forecasting performance against variants of the HU models across multiple forecast horizons. Our findings indicate that overall the HUts model not only provides more precise point forecasts but also shows robustness against data irregularities, such as those observed in countries with historical outliers. The integration of signature-based methods enables the HUts model to capture complex patterns in mortality data, making it a powerful tool for actuaries and demographers. Prediction intervals are also constructed with bootstrapping methods.

It is a joint work with Zhong Jing Yap, Dharini Pathmanathan (University of Malaya, Malaysia).

For more details, see [27].

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

Participants: Issam Moindjie, Sophie Dabo, Cristian Preda.

Multivariate functional data are 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 [21].

8.22 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 published in the CSDA journal [40].

8.23 Axis 3: Statistical learning with categorical functional data

Participants: Cristian Preda, Quentin Grimonprez.

We introduce unsupervised and supervised models for categorical functional data. Multiple correspondence analysis is extended to categorical functional data and principal components are used as latent variables for clustering and regression models. An application on the DAMAGE medical dataset is presented. A presentation of these models is done in [34].

8.24 Axis 4: Data-driven cluster analysis identifies distinct types of metabolic dysfunction-associated steatotic liver disease.

Participants: Guillemette Marot.

Metabolic dysfunction-associated steatotic liver disease (MASLD) exhibits considerable variability in clinical outcomes. Identifying specific phenotypic profiles within MASLD is essential for developing targeted therapeutic strategies. Here we investigated the heterogeneity of MASLD using partitioning around medoids clustering based on six simple clinical variables in a cohort of 1,389 individuals living with obesity. The identified clusters were applied across three independent MASLD cohorts with liver biopsy (totaling 1,099 participants), and in the UK Biobank to assess the incidence of chronic liver disease, cardiovascular disease and type 2 diabetes. Results unveiled two distinct types of MASLD associated with steatohepatitis on histology and liver imaging. The first cluster, liver-specific, was genetically linked and showed rapid progression of chronic liver disease but limited risk of cardiovascular disease. The second cluster, cardiometabolic, was primarily associated with dysglycemia and high levels of triglycerides, leading to a similar incidence of chronic liver disease but a higher risk of cardiovascular disease and type 2 diabetes. Analyses of samples from 831 individuals with available liver transcriptomics and 1,322 with available plasma metabolomics highlighted that these two types of MASLD exhibited distinct liver transcriptomic profiles and plasma metabolomic signatures, respectively. In conclusion, these data provide preliminary evidence of the existence of two distinct types of clinically relevant MASLD with similar liver phenotypes at baseline, but each with specific underlying biological profiles and different clinical trajectories, suggesting the need for tailored therapeutic strategies. This work was published in Nature Medicine [25].

8.25 Axis 4: Assessing the multi-dimensional effects of air pollution on maternal complications and birth outcomes : A structural equation modeling approach

Participants: Sophie Dabo-Niang.

This cross-sectional study aims to investigate the direct and indirect relationships between exposure to a metal mixture in air and adverse pregnancy outcomes across gestational stages. It reveals time-dependent associations between a metal mixture in $PM_{2.5}$ exposure and adverse pregnancy outcomes, highlights the need to address dust in $PM_{2.5}$, and provides additional evidence for understanding the pathway of the pollution effects on fetal health. 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 and Boubakari Ibrahimou (Florida International University, Miami, FL, USA) [18].

8.26 Axis 4: Demonstrating the relevance of spatial-functional statistical analysis in marine ecological studies : The case of environmental variations in micronektonic layers

Participants: Sophie Dabo-Niang.

In this study, we conducted an analysis of a multifrequency acoustics dataset acquired from scientific echosounders in the West African water. Our objective was to explore the spatial arrangement of marine organism aggregations. We investigated various attributes of these intricate biological entities, such as thickness, relative density, and depth, in relation to their surroundings. These environmental conditions were represented at a fine scale using a towed multiparameter system. This study is closely intertwined with two key domains: Fisheries acoustics techniques and functional data analysis. Fisheries acoustics techniques facilitate the collection of high-resolution spatial and temporal data concerning marine organisms at various depths and spatial scales, all without causing any disturbance. On the other hand, spatial-functional data analysis is a statistical approach for examining data characterised by functional attributes distributed across a spatial domain. This analysis encompasses dimension reduction techniques, as well as supervised and unsupervised methods, which take into consideration spatial dependencies within extensive dataset.

See for more details [19].

It is a joint work with Yoba Kande (IRD Senegal nd University of Lille), Ndague Diogoul, Patrice Brehmer and Yannick Perrot (IRD, Senegal) and Papa Ngom (University of Dakar, Senegal).

8.27 Axis 4: Aspects of the gender gap in Mathematics

Participants: Sophie Dabo-Niang.

As a contribution to some aspects of the gender gap in mathematics, this paper provides a detailed analysis of mathematicians answers to the Global Survey of Scientists. The questions we address are the following: are the answers from all sciences and the answers of mathematics or applied mathematics similar? In which cases is the situation in mathematics or in applied mathematics different from the situation in all sciences? Better or worse?

For more details, see [13]. It is a joint work with Maria Esteban (University of Paris Dauphine, France), Colette Guillopé (Université Paris-Est Créteil) and Marie-Françoise Roy (Université de Rennes I).

8.28 Axis 4: Functional data geometric morphometrics with machine learning for craniodental shape classification in shrews

Participants: Sophie Dabo-Niang.

This work proposes a functional data analysis approach for morphometrics in classifying three shrew species (S. murinus, C. monticola, and C. malayana) from Peninsular Malaysia. Functional data geometric morphometrics (FDGM) for 2D landmark data is introduced and its performance is compared with classical geometric morphometrics (GM). The FDGM approach converts 2D landmark data into continuous curves, which are then represented as linear combinations of basis functions. The landmark data was obtained from 89 crania of shrew specimens based on three craniodental views (dorsal, jaw, and lateral). Principal component analysis and linear discriminant analysis were applied to both GM and FDGM methods to classify the three shrew species. This study also compared four machine learning approaches (naïve Bayes, support vector machine, random forest, and generalised linear model) using predicted PC scores obtained from both methods (a combination of all three craniodental views and individual views). The analyses favoured FDGM and the dorsal view was the best view for distinguishing the three species.

It is a joint work with Aneesha Balachandran Pillay, Dharini Pathmanathan, Sophie Dabo-Niang, Arpah Abu and and Hasmahzaiti Omar from University of Malaya, Malaysia. See for more details [24].

8.29 Axis 4: A Novel Unstained Blood Smears Multispectral Images Normalization for automatic and rapid diagnosis of malaria

Participants: Sophie Dabo-Niang.

This work proposes a novel method for achieving this normalization, aiming to improve the accuracy and reliability of the diagnostic process. This method is based on estimating the Bright reference image, which captures the luminosity, and the contrast variability function from the background region of the image. This is achieved through two distinct resampling methodologies, namely Gaussian random field simulation by variogram analysis and Bootstrap resampling. A method for handling the intensity saturation issue of certain pixels is also proposed, which involves outlier imputation. Both of these proposed approaches for image normalization are demonstrated to outperform existing methods for multispectral and multimodal unstained blood smear images, as measured by the Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), Zero mean Sum of Absolute Differences (ZSAD), Peak Signal to Noise Ratio (PSNR), and Absolute Mean Brightness Error (AMBE). These methods not only improve the image contrast but also preserve its spectral footprint and natural appearance more accurately. The normalization technique employing Bootstrap resampling significantly reduces the acquisition time for multimodal and multispectral images by 66%. Moreover, the processing time for Bootstrap resampling is less than 4% of the processing time required for Gaussian random field simulation.

It is a joint work with Solange Doumun Oulai and Jérémie Zoueu from San Pedro Universuty (Ivory Coast). See for more details [15].

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

Diagrams Technologies startup

Participants: Christophe Biernacki, Cristian Preda.

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

9.2 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).

Axel Potier defended his PhD thesis on November 12 2024 [36].

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)

Silmach

Participants: Christophe Biernacki, Mustapha Atmani.

Through their joint ROAD-AI project, Inria and Cerema are jointly studying digital tools allowing these phenomena to be modeled using structural instrumentation. This initiative is complemented and reinforced by the SIRCAPASS project coordinated by the company SilMach and which aims to use new passive MEMS (Micro Electro-Mechanical Systems) sensor technology for this instrumentation.

In this context, Mustapha Atmani began his PhD thesis on December 1 2024 entitled “Statistical processing of “low data” from passive sensors: application to the monitoring of engineering structures”. The co-supervision is ensured by André Orcesi from Cerema.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Participation in other International Programs

FANE-MATH-PE

Participants: Sophie Dabo.

Title: France-Afrique Network for Mathematics of Sustainable Planet Earth. FANE-MATH-PE aims to develop cutting-edge applied mathematical tools for critical global challenges, including climate change, ocean resource management, and prevention and control of epidemics and cancers. This long-term project seeks to build a sustainable, diverse network of researchers from France, Africa, and beyond, uniting young talents and seasoned experts, women and men alike, in applied mathematics.

Partner Institution(s) : French institutions (CNRS, IRD, INSERM, University of Lille) and several prominent African institutions (AIMS centers - South Africa and Senegal, NRF, North West University, University of Johannesburg and University of Pretoria - South Africa, University of Carthage - Tunisia, Universities of Mohammed I and Mohamed V - Morocco, UCAD, Gaston-Berger and UADB - Senegal, Abomey Calavi-IMSP - Benin, Universities of Arba Minch and Hawassa - Ethiopia, Adama Science and Technology University - Ethiopia, San-Pedro University - Ivory Coast, University of NDjamena - Chad) and worldwide (Florida International University - USA, University College Dublin - Ireland, Saint Andrew University -Scotland, University of Manitoba - Canada).

Date/Duration : 2024 - 2027

10.2 International research visitors

10.2.1 Visits to international teams

Research stays abroad

Participants: Sophie Dabo.

Visited institution: CRM, University of Montréal and Fields Institute (Toronto)

Country: Canada

Dates : August 2024-February 2025

Context of the visit: Research visit

Mobility program/type of mobility: CNRS sabbatical

10.3 National initiatives

10.3.1 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).

10.3.2 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

10.3.3 ANR

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

10.3.4 FHU

A FHU is a federative project and a label necessary to postulate for a RHU.

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 Bililles 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 Bililles platform in WP3 for the research of a molecular signature predictive of the treatment response (resistance and complication).

10.3.5 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)

10.3.6 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.

10.3.7 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.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Participants: Christophe Biernacki.

General chair, scientific chair

- Christophe Biernacki co-organized at Sorbonne Université (Paris) on October 4 2024 the one-day workshop Frugalias dedicated to frugality in AI and Statistics. It gathered about 100 participants. He also gave an introductive keynote entitled "Historical perspective of frugality in AI and statistics" [32].

- Sophie Dabo, Cristian Preda and Hemant Tyagi organized at Inria Lille on 14 and 15 March 2024 the Workshop on functional data analysis fda-lille.

Member of the conference program committees Cristian Preda was member of The 25th Conference of the Romanian Society of Probability and Statistics conference program committee.

Member of the editorial boards

- Christophe Biernacki is an Associate Editor for the international journal Advances in Data Analysis and Classification (ADAC).

- Cristian Preda is an Associate Editor for Methodology and Computing in Applied Probability (MCAP).

Reviewer - reviewing activities

- Christophe Biernacki acted as a reviewer for different journals (Statistics and Computing,Journal of Classification, Computational Statistics and Data Analysis Journal of Computational and Graphical Statistics, Computational Statistics, The Austrian Journal of Statistics, Journal of Classification) and a conference (CAp 2024).

11.1.2 Invited talks

- Christophe Biernacki gave invited talks in France [28], Austria [30] and UK [29].
- Hemant Tyagi gave talks at Université Catholique de Louvain, NTU Singapore, One world seminar on mathematics of machine learning, University of Grenoble (GIPSA Lab).
- Cristian Preda gave talks to Journeée scientifique commune aux Fédérations NORMASTIC et Normandie-Mathématiques and also to the celebration of the 60th anniversary of the Romanian Academy Institute for Mathematical Statisics [34].

11.1.3 Leadership within the scientific community

- Christophe Biernacki was elected as the President of the SFdS (Société Française de Statistique) since July 2024, which is the French society specialized in Statistics, whose mission is to promote the use of statistics and its understanding and to foster its methodological developments.

11.1.4 Scientific expertise

- Christophe Biernacki has evaluated one ANR project and also one Phd Cifre ANRT project.

11.1.5 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".

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- 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

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

11.2.2 PhD supervision

- 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. He defended his PhD thesis on November 12 2024 [36].
- 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 co variates, from connected medical objects. Started in September 2023 under the supervision of Christophe Biernacki and Cristian Preda.
- Mustapha Atmani began his PhD thesis on December 1 2024 entitled "Statistical processing of "low data" from passive sensors: application to the monitoring of engineering structures". The co-supervision is ensured by André Orcesi from Cerema.
- François Bassac started his PhD thesis in October 2022 on models based on functional data or prediction of athlets sport performances. This is a PhD thesis in collaboration with the Decathlon company. This work is supervised by Cristian Preda.
- Hugo Cannafarina started his PhD thesis in September 2024 on selection variable techniques for survival models with recurrent events and competitive risks in high dimensional setting, in particular with omics data. That work is supervised by Guillemette Marot.

11.2.3 Juries

- Christophe Biernacki acted as a reviewer for 1 PhD thesis and for 1 HdR. He also participated to a jury at University of Nantes for recruiting an assistant professor.
- Hemant Tyagi was an external examiner for a PhD defense at the University of Warwick (Dept. of Mathematics) in January 2025.
- Cristian Preda acted as a reviewer for 2 phd theses (Bordeaux and Poitiers Universities).

11.3 Popularization

11.3.1 Participation in Live events

- Christophe Biernacki participated for one day to the "Fête de la science"' at Cité des Sciences (Paris).

12 Scientific production

12.1 Major publications

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

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

[3] C. Biernacki and A. Lourme. 'Unifying Data Units and Models in (Co-)Clustering'. In: *Advances in Data Analysis and Classification* 12.41 (May 2018). URL: https://hal.archives-ouvertes.fr/hal-01653881.

[4] A. Celisse. 'Optimal cross-validation in density estimation with the L2-loss'. In: *The Annals of Statistics* 42.5 (2014), pp. 1879–1910. URL: https://hal.archives-ouvertes.fr/hal-00337058.

[5] S. Dabo-Niang, C. Ternynck and A.-F. Yao. 'Nonparametric prediction in the multivariate spatial context'. In: *Journal of Nonparametric Statistics* 28.2 (2016), pp. 428–458. DOI: 10.1080/10485252.2016.01.007. URL: https://hal.inria.fr/hal-01425932.

[6] J. Dubois, V. Dubois, H. Dehondt, P. Mazrooei, C. Mazuy, A. A. Sérandour, C. Gheeraert, P. Guillaume, E. Baugé, B. Derudas, N. Hennuyer, R. Paumelle, G. Marot, J. S. Carroll, M. Lupien, B. Staels, P. Lefebvre and J. Eeckhoute. 'The logic of transcriptional regulator recruitment architecture at cis -regulatory modules controlling liver functions'. In: *Genome Research* 27.6 (June 2017), pp. 985–996. DOI: 10.1101/gr.217075.116. URL: https://hal.archives-ouvertes.fr/hal-01647846.

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

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

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

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

12.2 Publications of the year

International journals

[11] I. Si-Ahmed, L. Hamdad, C. J. Agonkoui, Y. Kande and S. Dabo-Niang. 'Principal component analysis of multivariate spatial functional data'. In: *Big Data Research* 39 (Feb. 2025), p. 100504. DOI: 10.1016/j.bdr.2024.100504. URL: https://inria.hal.science/hal-04905084 (cit. on p. 15).

[12] E. Araya, G. Braun and H. Tyagi. 'Seeded graph matching for the correlated Gaussian Wigner model via the projected power method'. In: *Journal of Machine Learning Research* 25.5 (24th Jan. 2024), 1–43. URL: https://hal.science/hal-03876872 (cit. on pp. 11, 12).

[13] S. Dabo-Niang, M. J. Esteban, C. Guillopé and M.-F. Roy. 'Aspects of the gender gap in Mathematics'. In: *European Mathematical Society Magazine* 131 (19th Mar. 2024), pp. 22–31. DOI: 10.4171/mag/178. URL: https://hal.science/hal-04278806 (cit. on p. 17).

[14] S. Dabo-Niang and C. Frévent. 'Uncovering Data Across Continua: An Introduction to Functional Data Analysis'. In: *Notices of the American Mathematical Society* 71 (1st Aug. 2024). DOI: 10.1090/noti2979. URL: https://inria.hal.science/hal-04905081 (cit. on p. 14).

[15] S. Doumun Oulai, S. Dabo-Niang and J. Zoueu. 'A Multispectral Blood Smear Background Images Reconstruction for Malaria Unstained Images Normalization'. In: *International Journal of Imaging Systems and Technology* 34.6 (4th Oct. 2024). DOI: 10.1002/ima.23182. URL: https://inria.hal.science/hal-04905075 (cit. on p. 18).

[16] P. A. L. Faye, E. Brunel, T. Claverie, S. M. Manou-Abi and S. Dabo-Niang. 'Automatic geomorphological mapping using ground truth data with coverage sampling and random forest algorithms'. In: *Earth Science Informatics* 17 (2024), pp. 3715–3732. DOI: 10.1007/s12145-024-01347-x. URL: https://hal.science/hal-04624799 (cit. on p. 14).

[17] C. Frévent, M.-s. Ahmed, S. Dabo-Niang and M. Genin. 'A Shared-Frailty Spatial Scan Statistic Model for Time-to-Event Data'. In: *Biometrical Journal* 66.5 (11th July 2024). DOI: 10.1002/bimj.202300200. URL: https://inria.hal.science/hal-04905079 (cit. on p. 13).

[18] B. Ibrahimou, N. Sun and S. Dabo-Niang. 'Assessing the multi-dimensional effects of air pollution on maternal complications and birth outcomes: A structural equation modeling approach'. In: *Hygiene and Environmental Health Advances* 12 (Dec. 2024), p. 100113. DOI: 10.1016/j.heha.2024.100113. URL: https://inria.hal.science/hal-04905076 (cit. on p. 16).

[19] Y. Kande, N. Diogoul, P. Brehmer, S. Dabo-Niang, P. Ngom and Y. Perrot. 'Demonstrating the relevance of spatial-functional statistical analysis in marine ecological studies : the case of environmental variations in micronektonic layers'. In: *Ecological Informatics* 81 (July 2024), p. 102547. DOI: 10.1016/j.ecoinf.2024.102547. URL: https://inria.hal.science/hal-04507879 (cit. on p. 17).

[20] T. H. Khoo, D. Pathmanathan, P. Otto and S. Dabo-Niang. 'A Markov-switching spatio-temporal ARCH model'. In: *Stat* 13.3 (15th July 2024). DOI: 10.1002/sta4.713. URL: https://inria.hal.science/hal-04905077 (cit. on p. 13).

[21] 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* (2024), p. 5. DOI: 10.1007/s11222-023-10324-1. URL: https://hal.science/hal-03908634 (cit. on p. 15).

[22] A. Picard-Weibel, G. Capson-Tojo, B. Guedj and R. Moscoviz. 'Bayesian Uncertainty Quantification for Anaerobic Digestion models'. In: *Bioresource Technology* 394 (Feb. 2024). DOI: 10.1016/j.biortech.2023.130147. URL: https://hal.science/hal-04592670.

[23] A. Picard-Weibel, G. Capson-Tojo, B. Guedj and R. Moscoviz. 'Bayesian uncertainty quantification for anaerobic digestion models'. In: *Bioresource Technology* 394 (Feb. 2024), p. 130147. DOI: 10.1016/j.biortech.2023.130147. URL: https://hal.inrae.fr/hal-04461829.

[24] A. B. Pillay, D. Pathmanathan, S. Dabo-Niang, A. Abu and H. Omar. 'Functional data geometric morphometrics with machine learning for craniodental shape classification in shrews'. In: *Scientific Reports* 14.1 (6th July 2024), p. 15579. DOI: 10.1038/s41598-024-66246-z. URL: https://inria.hal.science/hal-04905078 (cit. on p. 17).

[25] V. Raverdy, F. Tavaglione, E. Chatelain, G. Lassailly, A. de Vincentis, U. Vespasiani-Gentilucci, S. F. Qadri, R. Caiazzo, H. Verkindt, C. Saponaro, J. Pattou Kerr-Conte, G. Baud, C. Marciniak, M. Chetboun, N. Oukhouya-Daoud, S. Blanck, J. Vandel, L. Olsson, R. Chakaroun, V. Gnemmi, E. Leteurtre, P. Lefebvre, J. Haas, H. Yki-Järvinen, S. Francque, B. Staels, C. W. Le Roux, V. Tremaroli, P. Mathurin, G. Marot, S. Romeo and F. Pattou. 'Data-driven cluster analysis identifies distinct types of metabolic dysfunction-associated steatotic liver disease.' In: *Nature Medicine*. Nature Medicine 30.12 (14th Dec. 2024), pp. 3624–3633. DOI: 10.1038/s41591-024-03283-1. URL: https://hal.univ-lille.fr/hal-04876000 (cit. on p. 16).

[26] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, J. Josse and C. Biernacki. 'Model-based Clustering with Missing Not At Random Data'. In: *Statistics and Computing* (18th June 2024). DOI: 10.1007/s11222-024-10444-2. URL: https://hal.science/hal-03494674 (cit. on p. 8).

[27] Z. J. Yap, D. Pathmanathan and S. Dabo-Niang. 'Forecasting mortality rates with functional signatures'. In: *ASTIN Bulletin* (9th Jan. 2025), pp. 1–24. DOI: 10.1017/asb.2024.38. URL: https://inria.hal.science/hal-04905083 (cit. on p. 15).

Invited conferences

[28] C. Biernacki, J. Jacques and C. Keribin. 'Model Based Co-Clustering: High Dimension and Estimation Challenges'. In: RMR 2024 : Modèles statistiques pour des données dépendantes et applications. Rouen, France, 19th June 2024. URL: https://inria.hal.science/hal-04867840 (cit. on pp. 9, 26).

[29] C. Biernacki, J. Jacques and C. Keribin. 'Model Based Co-Clustering: High Dimension and Estimation Challenges'. In: CFE-CMStatistics 2024 - The 18th International Joint Conference on Computational and Financial Econometrics (CFE) and Computational and Methodological Statistics (CMStatistics). Londres, United Kingdom, 14th Dec. 2024. URL: https://inria.hal.science/hal-04867739 (cit. on pp. 8, 26).

[30] C. Biernacki and V. Vandewalle. 'An EM Stopping Rule for Avoiding Degeneracy in Gaussian-based Clustering with Missing Data'. In: *Advances in Intelligent Systems and Computing*. SMPS 2024 - International Conference on Soft Methods in Probability and Statistics. Vol. AISC-1458. Combining, Modelling and Analyzing Imprecision, Randomness and Dependence. Salsburg, Austria, 3rd Sept. 2024, pp. 27–35. DOI: 10.1007/978-3-031-65993-5_4. URL: https://inria.hal.science/hal-04867801 (cit. on pp. 9, 26).

International peer-reviewed conferences

[31] R. Adams, J. Shawe-Taylor and B. Guedj. 'Controlling Multiple Errors Simultaneously with a PAC-Bayes Bound'. In: NeurIPS 2024. Vancouver, Canada, Dec. 2024. URL: https://inria.hal.science/hal-03573458.

[32] C. Biernacki. 'MISE EN PERSPECTIVE HISTORIQUE DE LA FRUGALITÉ EN IA ET STATISTIQUE'. In: WORKSHOP FRUGALIAS (frugalité en intelligence artificielle et en statistique). Paris, France, 4th Oct. 2024. URL: https://inria.hal.science/hal-04868165 (cit. on p. 25).

National peer-reviewed Conferences

[33] 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 (cit. on p. 10).

Conferences without proceedings

[34] C. Preda and Q. Grimonprez. 'Statistical learning with categorical functional data: Application to pathways of elderly patients after hospitalization. The DAMAGE study.' In: Workshop Center of Mathematical Statistics 60. Bucharest, Romania, 15th Oct. 2024. URL: https://hal.science/hal-04917819 (cit. on pp. 16, 26).

Doctoral dissertations and habilitation theses

[35] E. Krönert. 'Anomaly detection in time series using breakpoint detection and multiple testing'. Université de Lille, 2nd Oct. 2024. URL: https://hal.science/tel-04891994 (cit. on p. 11).

[36] A. Potier. 'Estimating substitution for optimised replenishment with slow movers products'. Université de Lille, 12th Nov. 2024. URL: https://inria.hal.science/tel-04868026 (cit. on pp. 9, 19, 27).

Reports & preprints

[37] E. Araya, M. Cucuringu and H. Tyagi. *Dynamic angular synchronization under smoothness constraints*. 2024. DOI: 10.48550/arXiv.2406.04071. URL: https://hal.science/hal-04904627 (cit. on p. 13).

[38] E. Krönert, A. Célisse and D. Hattab. *FDR control for Online Anomaly Detection*. 16th Dec. 2024. URL: https://hal.science/hal-04321622 (cit. on p. 14).

[39] E. Krönert, D. Hattab and A. Celisse. *Breakpoint based online anomaly detection*. 5th Feb. 2024. URL: https://hal.science/hal-04440349 (cit. on p. 11).

[40] I.-A. Moindjié, C. Preda and S. Dabo-Niang. *Fusion regression methods with repeated functional data*. 20th Sept. 2024. URL: https://hal.science/hal-04176783 (cit. on p. 15).

[41] H. Tyagi. *Joint Learning of Linear Dynamical Systems under Smoothness Constraints*. 2024. DOI: 10.48550/arXiv.2406.01094. URL: https://hal.science/hal-04904614 (cit. on p. 12).

[42] H. Tyagi and D. Efimov. *Learning linear dynamical systems under convex constraints*. 15th Jan. 2024. URL: https://hal.science/hal-04394297 (cit. on p. 12).

Other scientific publications

[43] C. Boinay, C. Biernacki, C. Preda and F. Foyer. *Testing Abnormality of a Sequence of Graphs: Application to Cybersecurity*. 12th Jan. 2024. URL: https://inria.hal.science/hal-04868124 (cit. on p. 10).

[44] C. Keribin, C. Biernacki and J. Jacques. *Model Based Co-Clustering: High Dimension and Estimation Challenges*. 11th Mar. 2024. URL: https://inria.hal.science/hal-04862826 (cit. on p. 9).