RESEARCH CENTRE

Lille - Nord Europe

IN PARTNERSHIP WITH:

CNRS, Université de Lille

2020 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 Team: 2010 September 01, updated into Project-Team: 2012 January 01

Keywords

Computer sciences and digital sciences

A3.1.4. – Uncertain data

A3.2.3. – Inference

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.5. – Bayesian methods

A3.4.7. – Kernel methods

A5.2. – Data visualization

A5.9.2. – Estimation, modeling

A6.2.3. – Probabilistic methods

A6.2.4. – Statistical methods

A6.3.3. – Data processing

A9.2. – Machine learning

Other research topics and application domains

B2.2.3. – Cancer

B9.5.6. – Data science

B9.6.3. – Economy, Finance

B9.6.5. – Sociology

1 Team members, visitors, external collaborators

Research Scientists

- Christophe Biernacki [Team leader, Inria, Senior Researcher, team leader until Nov 2020, HDR]
- Benjamin Guedj [Inria, Researcher]
- Hemant Tyagi [Inria, Researcher]

Faculty Members

- Cristian Preda [Team leader, Université de Lille, Professor, team leader from Dec 2020, HDR]
- Vlad Barbu [Université de Rouen, Associate Professor, until Feb 2020, HDR]
- Alain Celisse [Université de Lille, Associate Professor, HDR]
- Sophie Dabo-Niang [Université de Lille, Professor, HDR]
- Philippe Heinrich [Université de Lille, Associate Professor]
- Serge Iovleff [Université de Lille, Associate Professor]
- Guillemette Marot [Université de Lille, Associate Professor, HDR]
- Vincent Vandewalle [Université de Lille, Associate Professor, HDR]

Post-Doctoral Fellows

- Florent Dewez [Inria]
- Vera Shalaeva [Inria, until Jun 2020]

PhD Students

- Reuben Adams [University College London, United Kingdom, from Sep 2020]
- Filippo Antonazzo [Inria]
- Yaroslav Averyanov [Inria]
- Felix Biggs [University College London, United Kingdom]
- Rajeev Bopche [Inria, from Oct 2020]
- Guillaume Braun [Insee]
- Wilfried Heyse [Inserm]
- Eglantine Karlé [Inria, from Nov 2020]
- Etienne Kronert [Wordline, from Jul 2020]
- Arthur Leroy [Université Paris-Descartes, until Sep 2020]
- Issam Ali Moindjie [Inria, from Oct 2020]
- Axel Potier [Inria, from Jul 2020]
- Antonin Schrab [University College London, United Kingdom, from Sep 2020]
- Antoine Vendeville [University College London, United Kingdom]
- Luxin Zhang [Wordline, CIFRE]

Technical Staff

- Maxime Brunin [Inria, Engineer, from Jul 2020]
- Iheb Eladib [Inria, Engineer, until Feb 2020]
- Quentin Grimonprez [Inria, Engineer, until Sep 2020]
- Etienne Kronert [Inria, Engineer, from Feb 2020 until Jun 2020]
- Issam Ali Moindjie [Inria, Engineer, until Sep 2020]
- Arthur Talpaert [Inria, Engineer, until Sep 2020]

Interns and Apprentices

- Theophile Cantelobre [Inria, from Feb 2020 until Jul 2020]
- Issa Dabo [Inria, from Jun 2020 until Aug 2020]
- Cadmos Kahale-Abdou [Inria, from Jul 2020 until Oct 2020]
- Komlan Midodzi Noukpoape [Inria, from Apr 2020 until Sep 2020]

Administrative Assistant

- Anne Rejl [Inria]

Visiting Scientist

- Apoorv Vikram Singh [Indian Institute of Science, Bangalore, India, until Jan 2020]

External Collaborators

- Jean-Francois Bouin [DiagRAMS Technologies, until Mar 2020]
- Margot Correard [DiagRAMS Technologies, until Mar 2020]

2 Overall objectives

2.1 Context

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

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

2.2 Goals

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

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

3 Research program

3.1 Research axis 1: Unsupervised learning

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

3.2 Research axis 2: Performance assessment

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

3.3 Research axis 3: Functional data

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

and applications to various domains, such as principal component analysis, clustering, regression and prediction.

3.4 Research axis 4: Applications motivating research

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

4 Application domains

4.1 Economic world

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

4.2 Biology

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

5 Highlights of the year

- Christophe Biernacki is now Deputy Scientific Director at Inria in charge of the national scientific domain “applied mathematics, computation and simulation”.
- Christophe Biernacki has been president of the scientific comitee of the JdS 2020.
- Benjamin Guedj has led the emerging Inria London Programme since 2019 and was appointed Scientific Director of the programme in September 2020. The partnernship involves Inria and University College London (United Kingdom) as of February 1, 2021 and the official kickoff.
- Sophie Dabo-Niang has been nominated in 2020 member of the Committee of Diversity of International Mathematical Union (IMU).
- DiagRAMS Technologies, a software editor dedicated to predictive maintenance, has been created this year. This start-up relies on the research of the MODAL team, developing a data analysis solution to anticipate breakdowns and malfunctions on industrial equipment.
- Cristian Preda is the new head of the MODAL team since December 2020. Vincent Vandewalle is the deputy director of the team.

5.1 Awards

Wilfried Heyse has been awarded at Spring of Cardiology prize for the best oral presentation of his poster [79].

Benjamin Guedj has obtained a best reviewer award (top 10% of the reviewers) for NeurIPS 2020.

Benjamin Guedj has co-authored a paper at NeurIPS 2020 which was selected for an oral presentation (top 3%) [39].

6 New software and platforms

6.1 New software

6.1.1 pycobra

Keywords: Statistics, Data visualization, Machine learning

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

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

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

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

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

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

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

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

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

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

URL: https://github.com/bhargavvader/pycobra

Publication: hal-01514059

Contact: Benjamin Guedj

Participants: Bhargav Srinivasa Desikan, Benjamin Guedj

6.1.2 MixtComp.V4

Keywords: Clustering, Statistics, Missing data, Mixed data

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

MixtComp is used as an R package, but its internals are coded in C++ using state of the art libraries for faster computation.

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

Contact: Christophe Biernacki

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

Partners: Université de Lille, CNRS

6.1.3 MASSICCC

Name: Massive Clustering with Cloud Computing

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

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

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

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

Contact: Christophe Biernacki

6.1.4 cfda

Name: Categorical functional data analysis

Keyword: Functional data

Functional Description: The R package cfda performs: - descriptive statistics for categorical functional data - dimension reduction end optimal encoding of states (correspondance multiple analyses towards functional data)

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

Contact: Cristian Preda

Participants: Cristian Preda, Quentin Grimonprez, Vincent Vandewalle

Partner: Université de Lille

6.1.5 PyRotor

Name: Python Route Trajectory Optimiser

Keywords: Optimization, Machine learning, Trajectory Modeling

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

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

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

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

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

Publication: hal-03024720

Contact: Florent Dewez

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

6.2 New platforms

6.2.1 MASSICCC Platform

MASSICCC is a demonstration platform giving access through a SaaS (service as a software) concept to data analysis libraries developed at Inria. It allows obtaining results either directly through a website specific display (specific and interactive visual outputs) or through an R data object download. It started in October 2015 for two years and is common to the Modal team (Inria Lille) and the Select team (Inria Saclay). In 2016, two packages have been integrated: Mixmod and MixtComp (see the specific section about MixtComp). In 2017, the BlockCluster package has been integrated and also a particular attention to provide meaningful graphical outputs (for Mixmod, MixtComp and BlockCluster) directly in the web platform itself has led to some specific developments. In 2019, a new version of the MixtComp software has been developed. From 2020, Julien Vandaele joined the MODAL team as a research engineer for upgrading both the MixtComp software and the MASSICCC platform.

7 New results

7.1 Axis 1: Model-based Co-clustering for Ordinal Data of Different Dimensions

Participants Christophe Biernacki.

This work has been motivated by a psychological survey on women affected by a breast tumor. Patients replied at different moments of their treatment to questionnaires with answers on ordinal scale. The questions relate to aspects of their life called dimensions. To assist the psychologists in analyzing the results, it is useful to emphasize a structure in the dataset. The clustering method achieves that by creating groups of individuals that are depicted by a representative of the group. From a psychological

position, it is also useful to observe how questions may be grouped. This is why a clustering should also be performed on the features, which is called a co-clustering problem. However, gathering questions that are not related to the same dimension does not make sense from a psychologist stance. Therefore, the present work corresponds to perform a constrained co-clustering method aiming to prevent questions from different dimensions from getting assembled in a same column-cluster. In addition, evolution of co-clusters along time has been investigated. The method relies on a constrained Latent Block Model embedding a probability distribution for ordinal data. Parameter estimation relies on a Stochastic EM-algorithm associated to a Gibbs sampler, and the ICL-BIC criterion is used for selecting the numbers of co-clusters. The resulting work has been accepted in an international journal in 2019 and the related R package ordinalClust has been accepted this year in another international journal [28].

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

7.2 Axis 1: Model-based Co-clustering for Mixed Type Data

Participants Christophe Biernacki.

Over decades, a lot of studies have shown the importance of clustering to emphasize groups of observations. More recently, due to the emergence of high-dimensional datasets with a huge number of features, co-clustering techniques have emerged and proposed several methods for simultaneously producing groups of observations and features. By synthesizing the dataset in blocks (the crossing of a row-cluster and a column-cluster), this technique can sometimes summarize better the data and its inherent structure. The Latent Block Model (LBM) is a well-known method for performing a co-clustering. However, recently, contexts with features of different types (here called mixed type datasets) are becoming more common. Unfortunately, the LBM is not directly applicable on this kind of dataset. The present work extends the usual LBM to the so-called Multiple Latent Block Model (MLBM) which is able to handle mixed type datasets. The inference is done through a Stochastic EM-algorithm embedding a Gibbs sampler and model selection criterion is defined to choose the number of row and column clusters. This method was successfully used on simulated and real datasets. This work is now accepted in an international journal [27].

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

7.3 Axis 1: Relaxing the Identically Distributed Assumption in Gaussian Co-clustering for High Dimensional Data

Participants Christophe Biernacki.

A co-clustering model for continuous data that relaxes the identically distributed assumption within blocks of traditional co-clustering is presented. The proposed model, although allowing more flexibility, still maintains the very high degree of parsimony achieved by traditional co-clustering. A stochastic EM algorithm along with a Gibbs sampler is used for parameter estimation and an ICL criterion is used for model selection. Simulated and real datasets are used for illustration and comparison with traditional co-clustering. This work has been submitted to an international journal [65].

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

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

Participants Christophe Biernacki, Vincent Vandewalle.

A generic method is introduced to visualize in a Gaussian-like way, and onto R^2, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have a thorough overview of the potentially complex clustering result. An entropic measure allows us to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis. This work is now published in an international journal [15].

This is a joint work with Matthieu Marbac from ENSAI.

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

Participants Christophe Biernacki.

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

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.

7.6 Axis 1: Organized Co-clustering for Textual Data Synthesis

Participants Christophe Biernacki.

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

a Stochastic EM algorithm embedding a Gibbs sampling step and the Poisson distribution. A paper has now been accepted in an international journal [29] and also in a national conference with international audience [47].

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

7.7 Axis 1: Model-Based Co-clustering with Co-variables

Participants Serge Iovleff.

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

7.8 Axis 1: Predictive Clustering

Participants Christophe Biernacki, Vincent Vandewalle.

Many data, for instance in biostatistics, contain some sets of variables which permit evaluating unobserved traits of the subjects (e.g. we ask question about how many pizzas, hamburgers, chips etc. are eaten to know how healthy are the food habits of the subjects). Moreover, we often want to measure the relations between these unobserved traits and some target variables (e.g. obesity). Thus, a two-steps procedure is often used: first, a clustering of the observations is performed on the sets of variables related to the same topic; second, the predictive model is fitted by plugging the estimated partitions as covariates. Generally, the estimated partitions are not exactly equal to the true ones. We investigate the impact of these measurement errors on the estimators of the regression parameters, and we explain when this two-steps procedure is consistent. We also present a specific EM algorithm which simultaneously estimates the parameters of the clustering and predictive models. This has led to the preprint [71] now submitted to an international journal.

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

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

Participants Filippo Antonazzo, Christophe Biernacki.

Clustering is impacted by the regular increase of sample sizes which provides opportunity to reveal information previously out of scope. However, the volume of data leads to some issues related to the need of many computational resources and also to high energy consumption. Resorting to binned data depending on an adaptive grid is expected to give proper answer to such green computing issues while not harming the quality of the related estimation. After a brief review of existing methods, a first application in the context of univariate model-based clustering is provided, with a numerical illustration of its advantages. Finally, an initial formalization of the multivariate extension is done, highlighting both issues and possible strategies. This work has been accepted to a national conference with international audience [43] and also to an international conference [33].

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

7.10 Axis 1: A Bumpy Journey: Exploring Deep Gaussian Mixture Models

Participants Christophe Biernacki.

The deep Gaussian mixture model (DGMM) is a framework directly inspired by the finite mixture of factor analysers model (MFA) and the deep learning architecture composed of multiple layers. The MFA is a generative model that considers a data point as arising from a latent variable (termed the score) which is sampled from a standard multivariate Gaussian distribution and then transformed linearly. The linear transformation matrix (termed the loading matrix) is specific to a component in the finite mixture. The DGMM consists of stacking MFA layers, in the sense that the latent scores are no longer assumed to be drawn from a standard Gaussian, but rather are drawn from a mixture of factor analysers model. Thus the latent scores are at one point considered to be the input of an MFA and also to have latent scores themselves. The latent scores of the DGMM's last layer only are considered to be drawn from a standard multivariate Gaussian distribution. In recent years, the DGMM has gained prominence in the literature: intuitively, this model should be able to capture complex distributions more precisely than a simple Gaussian mixture model. We show in this work that while the DGMM is an original and novel idea, in certain cases it is challenging to infer its parameters. In addition, we give some insights to the probable reasons of this difficulty. Experimental results are provided on github: https://github.com/ansubmissions/ICBINB, alongside an R package that implements the algorithm and a number of ready-to-run R scripts. This work has been accepted in an international workshop [46].

This is a joint work with Margot Selosse (PhD student) and Julien Jacques, both from Université de Lyon 2, and also Isobel Claire Gormley from University College Dubin (Ireland).

7.11 Axis 1: Multiple partition clustering subspaces

Participants Vincent Vandewalle.

In model based clustering, it is often supposed that only one clustering latent variable explains the heterogeneity of the whole dataset. However, in many cases several latent variables could explain the heterogeneity of the data at hand. Finding such class variables could result in a richer interpretation of the data. In the continuous data setting, a multi-partition model based clustering is proposed. It assumes the existence of several latent clustering variables, each one explaining the heterogeneity of the data with respect to some clustering subspace. It allows to simultaneously find the multi-partitions and the related subspaces. Parameters of the model are estimated through an EM algorithm relying on a probabilistic reinterpretation of the factorial discriminant analysis. A model choice strategy relying on the BIC criterion is proposed to select to number of subspaces and the number of clusters by subspace. The obtained results are thus several projections of the data, each one conveying its own clustering of the data.

This work in now published in [31].

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

Participants Hemant Tyagi.

Given a measurement graph $G = ([n], E)$ and an unknown signal $r \in R^n$, we investigate algorithms for recovering r from pairwise measurements of the form $r_i - r_j; \{i, j\} \in E$. This problem arises in a variety of applications, such as ranking teams in sports data and time synchronization of distributed networks. Framed in the context of ranking, the task is to recover the ranking of n teams (induced by r) given a

small subset of noisy pairwise rank offsets. We propose a simple SVD-based algorithmic pipeline for both the problem of time synchronization and ranking. We provide a detailed theoretical analysis in terms of robustness against both sampling sparsity and noise perturbations with outliers, using results from matrix perturbation and random matrix theory. Our theoretical findings are complemented by a detailed set of numerical experiments on both synthetic and real data, showcasing the competitiveness of our proposed algorithms with other state-of-the-art methods.

This is joint work with Alexandre d'Aspremont (CNRS & ENS, Paris) and Mihai Cucuringu (University of Oxford, United Kingdom) and has now been published in an international journal [19].

7.13 Axis 1: Regularized spectral methods for clustering signed networks

Participants Hemant Tyagi.

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

This is joint work with Mihai Cucuringu (University of Oxford, United Kingdom), Apoorv Vikram Singh (NYU), Deborah Sulem (University of Oxford, United Kingdom). It was initiated when Apoorv Vikram Singh visited the MODAL team to work with Hemant Tyagi from Oct 2019-Jan 2020. It is currently under review in an international journal. A summary of the results was presented at the GCLR (Graphs and more Complex structures for Learning and Reasoning) workshop at AAAI 2021 (https://sites.google.com/view/gclr2021/accepted-papers).

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

Participants Hemant Tyagi.

Given an undirected measurement graph $G = ([n], E)$, the classical angular synchronization problem consists of recovering unknown angles $\theta_1, \ldots, \theta_n$ from a collection of noisy pairwise measurements of the form $(\theta_i - \theta_j) \bmod 2\pi$, for each $\{i, j\} \in E$. This problem arises in a variety of applications, including computer vision, time synchronization of distributed networks, and ranking from preference relationships. In this paper, we consider a generalization to the setting where there exist k unknown groups of angles $\theta_{l,1}, \ldots, \theta_{l,n}$, for $l = 1, \ldots, k$. For each $\{i, j\} \in E$, we are given noisy pairwise measurements of the form $\theta_{\ell,i} - \theta_{\ell,j}$ for an *unknown* $\ell \in \{1, 2, \ldots, k\}$. This can be thought of as a natural extension of the

angular synchronization problem to the heterogeneous setting of multiple groups of angles, where the measurement graph has an unknown edge-disjoint decomposition $G = G_1 \cup G_2 \ldots \cup G_k$, where the G_i's denote the subgraphs of edges corresponding to each group. We propose a probabilistic generative model for this problem, along with a spectral algorithm for which we provide a detailed theoretical analysis in terms of robustness against both sampling sparsity and noise. The theoretical findings are complemented by a comprehensive set of numerical experiments, showcasing the efficacy of our algorithm under various parameter regimes. Finally, we consider an application of bi-synchronization to the graph realization problem, and provide along the way an iterative graph disentangling procedure that uncovers the subgraphs G_i, $i = 1, \ldots, k$ which is of independent interest, as it is shown to improve the final recovery accuracy across all the experiments considered.

This is joint work with Mihai Cucuringu (University of Oxford, United Kingdom) and is currently under review in an international journal.

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

Participants Christophe Biernacki, Guillaume Braun, Hemant Tyagi.

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

This PhD work has recently begun and has let to a national conference paper with international audience [34]. An extended version has been submitted and accepted to an international conference for 2021.

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

Participants Hemant Tyagi.

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

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

This is joint work with Michael Fanuel (KU Leuven). It is currently under review in an international journal and is undergoing revision.

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

Participants Hemant Tyagi.

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

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

This is currently under review in an international journal, and is undergoing revision.

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

Participants Hemant Tyagi.

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

This is joint work with Stephane Chretien (National Physical Laboratory, United Kingdom & Alan Turing Institute, London) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It has now been published in an international journal [17].

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

Participants Hemant Tyagi.

Consider an unknown smooth function $f : [0,1]^d \to R$, and assume we are given n noisy mod 1 samples of f, i.e. $y_i = (f(x_i) + \eta_i) \bmod 1$, for $x_i \in [0,1]^d$, where η_i denotes the noise. Given the samples $(x_i, y_i)_{i=1}^n$, our goal is to recover smooth, robust estimates of the clean samples $f(x_i) \bmod 1$. We formulate a natural approach for solving this problem, which works with angular embeddings of the noisy mod 1 samples over

the unit circle, inspired by the angular synchronization framework. This amounts to solving a smoothness regularized least-squares problem – a quadratically constrained quadratic program (QCQP) – where the variables are constrained to lie on the unit circle. Our proposed approach is based on solving its relaxation, which is a *trust-region sub-problem* and hence solvable efficiently. We provide theoretical guarantees demonstrating its robustness to noise for adversarial, as well as random Gaussian and Bernoulli noise models. To the best of our knowledge, these are the first such theoretical results for this problem. We demonstrate the robustness and efficiency of our proposed approach via extensive numerical simulations on synthetic data, along with a simple least-squares based solution for the unwrapping stage, that recovers the original samples of f (up to a global shift). It is shown to perform well at high levels of noise, when taking as input the denoised modulo 1 samples. Finally, we also consider two other approaches for denoising the modulo 1 samples that leverage tools from Riemannian optimization on manifolds, including a Burer-Monteiro approach for a semidefinite programming relaxation of our formulation. For the two-dimensional version of the problem, which has applications in synthetic aperture radar interferometry (InSAR), we are able to solve instances of real-world data with a million sample points in under 10 seconds, on a personal laptop.

This is joint work with Mihai Cucuringu (University of Oxford, United Kingdom) and was mostly done while Hemant Tyagi was affiliated to the Alan Turing Institute. It has now been published in an international journal [18].

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

Participants Pascal Germain.

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

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

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

7.21 Axis 2: Improved PAC-Bayesian Bounds for Linear Regression

Participants Pascal Germain, Vera Shalaeva.

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

It is a joint work with Mihaly Petreczky and Alireza Fakhrizadeh Esfahani from Université de Lille. It has been accepted for publication as part of the AAAI 2020 conference [41].

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

Participants Pascal Germain.

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

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

7.23 Axis 2: PAC-Bayes and Domain Adaptation

Participants Pascal Germain.

In machine learning, Domain Adaptation (DA) arises when the distribution generating the test (target) data differs from the one generating the learning (source) data. It is well known that DA is a hard task even under strong assumptions, among which the covariate-shift where the source and target distributions diverge only in their marginals, i.e. they have the same labeling function. Another popular approach is to consider a hypothesis class that moves closer the two distributions while implying a low-error for both tasks. This is a VC-dim approach that restricts the complexity of a hypothesis class in order to get good generalization. Instead, we propose a PAC-Bayesian approach that seeks for suitable weights to be given to each hypothesis in order to build a majority vote. We prove a new DA bound in the PAC-Bayesian context. This leads us to design the first DA-PAC-Bayesian algorithm based on the minimization of the proposed bound. Doing so, we seek for a ρ-weighted majority vote that takes into account a trade-off between three quantities. The first two quantities being, as usual in the PAC-Bayesian approach, (a) the complexity of the majority vote (measured by a Kullback-Leibler divergence) and (b) its empirical risk (measured by the ρ-average errors on the source sample). The third quantity is (c) the capacity of the majority vote to distinguish some structural difference between the source and target samples.

This work has been published in the journal Neurocomputing [24].

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

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

Participants Pascal Germain, Paul Viallard.

We propose a PAC-Bayesian theoretical study of the two-phase learning procedure of a neural network introduced by Kawaguchi et al. [84]. In this procedure, a network is expressed as a weighted combination of all the paths of the network (from the input layer to the output one), that we reformulate as a PAC-Bayesian majority vote. Starting from this observation, their learning procedure consists in (1) learning "prior" network for fixing some parameters, then (2) learning a "posterior" network by only allowing a modification of the weights over the paths of the prior network. This allows us to derive a PAC-Bayesian

generalization bound that involves the empirical individual risks of the paths (known as the Gibbs risk) and the empirical diversity between pairs of paths. Note that similarly to classical PAC-Bayesian bounds, our result involves a KL-divergence term between a "prior" network and the "posterior" network. We show that this term is computable by dynamic programming without assuming any distribution on the network weights.

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

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

7.25 Axis 2: PAC-Bayesian Bound for the Conditional Value at Risk

Participant: Benjamin Guedj

Participants Benjamin Guedj.

Conditional Value at Risk (CVaR) is a family of "coherent risk measures" which generalize the traditional mathematical expectation. Widely used in mathematical finance, it is garnering increasing interest in machine learning, e.g. as an alternate approach to regularization, and as a means for ensuring fairness. This paper presents a generalization bound for learning algorithms that minimize the CVaR of the empirical loss. The bound is of PAC-Bayesian type and is guaranteed to be small when the empirical CVaR is small. We achieve this by reducing the problem of estimating CVaR to that of merely estimating an expectation. This then enables us, as a by-product, to obtain concentration inequalities for CVaR even when the random variable in question is unbounded.

Joint work with Mhammedi Zakaria (Australian National University) and Robert Williamson. Published: [39]

7.26 Axis 2: PAC-Bayesian Contrastive Unsupervised Representation Learning

Participants Benjamin Guedj, Pascal Germain.

Contrastive unsupervised representation learning (CURL) is the state-of-the-art technique to learn representations (as a set of features) from unlabelled data. While CURL has collected several empirical successes recently, theoretical understanding of its performance was still missing. In a recent work, Arora et al. [86] provide the first generalisation bounds for CURL, relying on a Rademacher complexity. We extend their framework to the flexible PAC-Bayes setting, allowing to deal with the non-iid setting. We present PAC-Bayesian generalisation bounds for CURL, which are then used to derive a new representation learning algorithm. Numerical experiments on real-life datasets illustrate that our algorithm achieves competitive accuracy, and yields generalisation bounds with non-vacuous values.

Joint work with Kento Nozawa (University of Tokyo & RIKEN, Japan). Published: [40]

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

Participants Benjamin Guedj.

This work studies clustering for possibly high dimensional data (e.g. images, time series, gene expression data, and many other settings), and rephrase it as low rank matrix estimation in the PAC-Bayesian framework. Our approach leverages the well known Burer-Monteiro factorisation strategy from large scale optimisation, in the context of low rank estimation. Moreover, our Burer-Monteiro factors are shown

to lie on a Stiefel manifold. We propose a new generalized Bayesian estimator for this problem and prove novel prediction bounds for clustering. We also devise a componentwise Langevin sampler on the Stiefel manifold to compute this estimator.

Joint work with Stéphane Chrétien (Université Lyon 2). Published: [35]

7.28 Axis 2:Kernel-Based Ensemble Learning in Python

Participants Benjamin Guedj.

We propose a new supervised learning algorithm for classification and regression problems where two or more preliminary predictors are available. We introduce KernelCobra, a non-linear learning strategy for combining an arbitrary number of initial predictors. KernelCobra builds on the COBRA algorithm which combined estimators based on a notion of proximity of predictions on the training data. While the COBRA algorithm used a binary threshold to declare which training data were close and to be used, we generalise this idea by using a kernel to better encapsulate the proximity information. Such a smoothing kernel provides more representative weights to each of the training points which are used to build the aggregate and final predictor, and KernelCobra systematically outperforms the COBRA algorithm. While COBRA is intended for regression, KernelCobra deals with classification and regression. KernelCobra is included as part of the open source Python package Pycobra (0.2.4 and onward). Numerical experiments were undertaken to assess the performance (in terms of pure prediction and computational complexity) of KernelCobra on real-life and synthetic datasets.

Published: [25]

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

Participants Benjamin Guedj.

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

Joint work with Juliette Rengot, Ecole de Ponts, ParisTech.

Published: [37]

7.30 Axis 2: Multiple change-points detection with reproducing kernels

Participants Alain Celisse.

We tackle the change-point problem with data belonging to a general set. We build a penalty for choosing the number of change-points in the kernel-based method of Harchaoui and Cappé [83]. This penalty generalizes the one proposed by Lebarbier [85] for a one-dimensional signal changing only through its mean. We prove a non-asymptotic oracle inequality for the proposed method, thanks to a new concentration result for some function of Hilbert-space valued random variables. Experiments on synthetic and real data illustrate the accuracy of our method, showing that it can detect changes in the whole distribution of data, even when the mean and variance are constant.

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

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

Participants Alain Celisse.

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

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

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

Participants Sophie Dabo-Niang.

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

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

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

Participants Sophie Dabo-Niang.

In order to identify mathematical modeling (including functional data analysis) and interdisciplinary research issues in evolutionary biology, epidemiology, epistemology, environmental and social sciences encountered by researchers in Mayotte, the first international conference on mathematical modeling (CIMOM'18) was held in Dembéni, Mayotte, from November 15 to 17, 2018, at the Centre Universitaire de Formation et de Recherche. The objective was to focus on mathematical research with interdisciplinarity. This contribution is a book discusses key aspects of recent developments in applied mathematical analysis and modeling. It was written after the international conference on mathematical modeling in Mayotte, where a call for chapters of the book was made. They were written in the form of journal articles, with new results extending the talks given during the conference and were reviewed by independent reviewers and book publishers It highlights a wide range of applications in the fields of biological and environmental sciences, epidemiology and social perspectives. Each chapter examines selected research

problems and presents a balanced mix of theory and applications on some selected topics. Particular emphasis is placed on presenting the fundamental developments in mathematical analysis and modeling and highlighting the latest developments in different fields of probability and statistics. The chapters are presented independently and contain enough references to allow the reader to explore the various topics presented.

It is a joint work with Solym Manou-Abi and Jean-Jacques Salone (Centre Universitaire de Mayotte). This book is to appear at Wiley (ISTE)

7.34 Axis 3: Categorical functional data analysis

Participants Cristian Preda, Quentin Grimonprez, Vincent Vandewalle.

The research on functional data analysis is very actual. The R package "fda" is the most famous one implementing methodology for functional data. To the best of our knowledge, and quite surprisingly, there is no recent researches devoted to categorical functional data despite its ability to model real situations in different fields of applications: health and medicine (status of a patient over time), economy (status of the market), sociology (evolution of social status), and so on. We have developed the methodology to visualize, do dimension reduction and extract feature from categorical functional data. For this, the *cfda* R package has been developed. This has led to the preprint [72] that will be submitted in an international journal.

7.35 Axis 3: Scan Statistics

Participants Cristian Preda, Alexandru Amarioarei.

The one dimensional discrete scan statistic is considered over sequences of random variables generated by block factor dependence models. Viewed as a maximum of an 1-dependent stationary sequence, the scan statistics distribution is approximated with accuracy and sharp bounds are provided. The longest increasing run statistics is related to the scan statistics and its distribution is studied. The moving average process is a particular case of block factor and the distribution of the associated scan statistics is approximated. Numerical results are presented.

This work was presented [44] and published: [26].

7.36 Axis 3: Clustering categorical functional data

Participants Cristian Preda, Vincent Vandewalle, Vlad Stefan Barbu.

The objective of this research direction was: (i) to propose possible modelling approaches of categorical functional data and (ii) to investigate the identifiability problem of such models. A first modelling framework is to consider that an observed functional data path represents a sample path of Markov process and thus n sample paths come from several, say K, different processes. Consequently, we have here a mixture of K different Markov processes. A second modelling framework is to consider that the observed sample path come from several semi-Markov processes. The parameter estimation is obtained through techniques based on the EM algorithm, while the selection of the number of classes is based on information criteria. An important problem is to determine the class membership for each sample paths, but our main concern that we have started to investigate is related to the identifiability problem. As far as we have studied, it seems that the identifiability of this type of models cannot be obtained in general, but only by imposing restrictions on the parameters of the model, cf. [81, 82]. Our work in progress is related to finding sufficiently general conditions that guarantee this identifiability.

7.37 Axis 3: Estimation of right-censored categorical functional data

Participants Cristian Preda, Vincent Vandewalle, Vlad Stefan Barbu.

As mentioned in Section 7.36, we are interested in modelling categorical functional data by means of semi-Markov processes. These processes generalize Markov processes, in the sense that the sojourn time in a state can be arbitrarily distributed, as opposed to the Markov case. For this reason, semi-Markov processes are flexible tools, more adapted to concrete applications as compared to Markov processes [80]. As in any modelling framework, it is clear that one crucial point is to obtain reliable estimators of the parameters of the model. A very important feature in many applications (e.g. survival analysis, reliability, etc.) is to take into account censored data. In the presence of right-censored sample paths, the estimation of semi-Markov processes in continuous time is still an open problem, while for discrete-time semi-Markov we have only an existing research in a non-parametric setting [87]. For this framework, we have already established the main setting, and derived the form of the Q-function for the EM algorithm. Several choices have to be made, that open different research paths: parametric versus non-parametric estimation for the sojourn time distributions, types of semi-Markov processes, considering mixtures for the sojourn time distributions, considering mixtures for the semi-Markov processes, etc. The next step of our work is to implement this estimation algorithm and to investigate, calibrate, adapt the algorithm. Another feature that we have not yet considered, but could be of great importance in some applications, is to investigate data that are censored under other censoring schemes, like censoring at the beginning of the sample path or interval censoring.

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

Participants Guillemette Marot, Vincent Vandewalle, Wilfried Heyse.

Since November 2019, Wilfried Heyse has started a PhD thesis granted by INSERM and supervised by Christophe Bauters, Guillemette Marot and Vincent Vandewalle. The aim is to identify earlier after myocardial infarction (MI) patients at high risk of developing left ventricular remodelling (LVR) that is quantified by imaging one year after MI or to identify patients with high risk of death. For that purpose, high throughput proteomic approach is used. This technology allows the measurement of 5000 proteins simultaneously. In parallel to these measures corresponding to the concentration of a protein in a plasma sample collected from one patient at a specific time, echocardiographic and clinical information have been collected on each of the 200 patients. One of the main challenge is to take into account the variations of the biomarkers according to the time (several measurement times), in order to improve the understanding of biological mechanisms involved on LVR or survival of the patient. Preliminary results have been presented in [38, 79].

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

7.39 Axis 4: Reject Inference Methods in Credit Scoring

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

The granting process of all credit institutions rejects applicants having a low credit score. Developing a scorecard, i.e. a correspondence table between a client's characteristics and his score, requires a learning dataset in which the target variable good/bad borrower is known. Rejected applicants are de facto excluded from the process. This biased learning population might have deep consequences on the scorecard relevance. Some works, mostly empirical ones, try to exploit rejected applicants in the scorecard building process. This work proposes a rational criterion to evaluate the quality of a scoring

model for the existing Reject Inference methods and dig out their implicit mathematical hypotheses. It is shown that, up to now, no such Reject Inference method can guarantee a better credit scorecard. These conclusions are illustrated on simulated and real data from the french branch of Crédit Agricole Consumer Finance (CACF). This has led to the preprint [63] which is now in revision in an international journal.

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

7.40 Axis 4: Usability study

Participants Vincent Vandewalle.

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

The discovery process can be modeled by a binary matrix, a matrix whose number of columns depends on the number of defects discovered by users. In this framework, they have proposed a probabilistic modeling of this matrix. They have included this modeling in a Bayesian framework where the number of problems and the probabilities of discovery are considered as random variables. In this framework, the article [32] as been published. It shows the interest of the approach compared to the approaches proposed in the state of the art in usability. The approach beyond point estimation also makes it possible to obtain the distribution of the number of problems and their respective probabilities given the discovery matrix.

The proposed model also allows to implement an approach aiming at measuring the value of additional information in relation to the discovery process. In this framework, they are writing a second paper and developing the R package useval available soon. This work has been presented in a conference [48].

This is a joint work with Alexandre Caron and Benoît Dervaux both from ULR 2694: METRICS.

7.41 Axis 4: Artificial intelligence for aviation

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

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

The article [21] is now published. It explains how, using flight recording data, it is possible to implement learning models on variables that have not been directly observed, and in particular to predict the drag and lift coefficients as a function of the angle and speed of the aircraft.

A second article is being to be submitted about the optimization of the aircraft's trajectory based on a consumption model learned from the data, and is available as a preprint [62]. The originality of the approach consists in decomposing the trajectory on a functional basis, and thus carrying out the optimization on the coefficients of the decomposition on this basis, rather than approaching the problem from the angle of optimal control. Furthermore, to guarantee compliance with aeronautical constraints, we have proposed an approach penalized by a deviation term from reference flights. A generic Python

module (PyRotor) to solve such optimization problems in conjunction with the proposed approach has been developed.

7.42 Axis 4: Domain Adaptation from a Pre-trained Source Model

Participants Christophe Biernacki, Pascal Germain, Luxin Zhang.

Traditional statistical learning paradigm assumes the consistency between train and test data distributions. This rarely holds in many real-life applications. The domain adaptation paradigm proposes a variety of techniques to overcome this issue. Most of the works in this area seek either for a latent space where source and target data share the same distribution, or for a transformation of the source distribution to match the target one. Both strategies require learning a model on the transformed source data. An original scenario is studied where one is given a model that has been constructed using expertise on the source data that is not accessible anymore. To use directly this model on target data, we propose to learn a transformation from the target domain to the source domain. Up to our knowledge, this is a new perspective on domain adaptation. This learning problem is introduced and formalized. We study the assumptions and the sufficient conditions mandatory to guarantee a good accuracy when using the source model directly on transformed target data. By pursuing this idea, a new domain adaptation method based on optimal transport is proposed. We experiment our method on a fraud detection problem. This work has been accepted to an international conference [42].

It is a joint work with Yacine Kessaci from Worldline company.

7.43 Other: Projection Under Pairwise Control

Participants Christophe Biernacki.

Visualization of high-dimensional and possibly complex (non-continuous for instance) data onto a low-dimensional space may be difficult. Several projection methods have been already proposed for displaying such high-dimensional structures on a lower-dimensional space, but the information lost is not always easy to use. Here, a new projection paradigm is presented to describe a non-linear projection method that takes into account the projection quality of each projected point in the reduced space, this quality being directly available in the same scale as this reduced space. More specifically, this novel method allows a straightforward visualization data in R2 with a simple reading of the approximation quality, and provides then a novel variant of dimensionality reduction. This work has now been accepted in an international journal [13].

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

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

Participants Christophe Biernacki.

We revisit the concept of sphere of gravitational activity, to which we give both a geometrical and physical meaning. This study aims to refine this concept in a much broader context that could, for instance, be applied to exo-planetary problems (in a Galactic stellar disc-StarPlanets system) to define a first order “border” of a planetary system. The methods used in this paper rely on classical Celestial Mechanics and develop the equations of motion in the framework of the 3-body problem (e.g. Star-Planet-Satellite System. We start with the basic definition of planet’s sphere of activity as the region of space in which it is feasible to assume a planet as the central body and the Sun as the perturbing

body when computing perturbations of the satellite's motion. We then investigate the geometrical properties and physical meaning of the ratios of Solar accelerations (central and perturbing) and planetary accelerations (central and perturbing), and the boundaries they define. We clearly distinguish throughout the paper between the sphere of activity, the Chebotarev sphere (a particular case of the sphere of activity), Laplace sphere, and the Hill sphere. The last two are often wrongfully thought to be one and the same. Furthermore, taking a closer look and comparing the ratio of the star's accelerations (central/perturbing) to that of the planetary acceleration (central/perturbing) as a function of the planeto-centric distance, we have identified different dynamical regimes which are presented in the semi-analytical analysis. This work has been published in an international journal [30].

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

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

COLAS company

Participants Christophe Biernacki.

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

PAY-BACK company

Participants Christophe Biernacki.

PAY-BACK Group is an audit firm specializing in the analysis and reliability of transactions. This bilateral contract aims at predicting store sales both from past sales (times series) and also by exploiting external covariates (of different types).

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

8.2 Bilateral grants with industry

Worldline

Participants Christophe Biernacki.

Worldline is the new world-class leader in the payments and transactional services industry, with a global reach. A PhD began in Feb. 2019 with Luxing Gang under the supervision of Christophe Biernacki, Pascal Germain (Laval University, Canada) and Yacine Kessaci (Worldline) on the topic of the domain adaptation from a pre-trained source model (with application to fraud detection in electronic payments).

ADEO

Participants Christophe Biernacki, Vincent Vandewalle.

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

EIT-Sysbooster: Nokia - Apsys/Airbus

Participants Alain Celisse.

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

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Inria International Labs

6PAC

Participants Benjamin Guedj *(coordinator)*.

- **Title:** *Making Probably Approximately Correct Learning Active, Sequential, Structure-aware, Efficient, Ideal and Safe*
- **Duration:** 2018–2022
- **Partners:** Machine Learning Group, CWI (The Netherlands)
- **Summary:** This project roots in statistical learning theory, which can be viewed as the theoretical foundations of machine learning. The most common framework is a setup in which one is given n training examples, and the goal is to build a predictor that would be efficient on new (similar) data. This efficiency should be supported by PAC (Probably Approximately Correct) guarantees, e.g. upper bounds on the excess risk of a predictor that hold with high probability. Such guarantees however often hold under stringent assumptions which are typically never met in real-life application, e.g. independent, identically distributed data. More realistic modelling of data has triggered many research efforts in several directions: first, accommodating possible data (e.g. dependent, heavy-tailed), and second, in the direction of sequential learning, in which the predictor can be built on the fly, while new data is gathered. We believe that an ever more realistic paradigm is active learning, a setup in which the learner actively requests data (possibly facing constraints, such as storage, velocity, cost, etc.) and adapts its queries to optimize its performance. The 3-years objective of 6PAC (where 6 stands for Sequential, Active, Efficient, Structured, Ideal, Safe — the six research directions we intend to contribute to) is to pave the way to new PAC generalization and sample-complexity upper and lower bounds beyond batch learning. Our ambition is to contribute to several learning setups, ranging from sequential learning (where data streams are collected) to adaptive and active learning (where data streams are requested by the learning algorithm).

9.1.2 Inria international partners

Benjamin Guedj leads The Inria London Programme, an initiative from Inria to increase the volume of scientific collaborations with the UK and in particular with the London region, with the prime partnership with University College London (United Kingdom).

More details at https://london.inria.fr

9.2 International research visitors

9.2.1 Visits of international scientists

- Apoorv Vikram Singh (IISc Bangalore, India) visited Hemant Tyagi from Oct 2019 to Jan 2020 to work on a project related to clustering of signed networks. This was partially funded by the Turing Institute, London. Apoorv worked under the joint supervision of Hemant Tyagi and Mihai Cucuringu (University of Oxford, United Kingdom) during this period.
- Déborah Sulem (PhD student, University of Oxford, United Kingdom) visited Hemant Tyagi on January 13–15, 2020.

9.3 European initiatives

9.3.1 FP7 & H2020 Projects

H2020 FAIR

Participants Guillemette Marot.

- **Acronym:** FAIR
- **Project title:** Flagellin aerosol therapy as an immunomodulatory adjunct to the antibiotic treatment of drug-resistant bacterial pneunomia
- **Coordinator:** JC. Sirard (Inserm, CIIL)
- **Duration:** 4 years (2020–2023)
- **Partners:** Inserm, Université de Lille, Free University of Berlin (Germany), Epithelix (Switzerland), Aerogen (Ireland), Statens Serum Institute (Denmark), CHRU Tours, Academic Medical Center of the University of Amsterdam (The Netherlands), University of Southampton (United Kingdom), European Respiratory Society (Switzerland)
- **Abstract:** The FAIR project aims at evaluating an alternative adjunct strategy to standard of care antibiotics for treating pneumonia caused by antibiotic-resistant bacteria: activation of the innate immune system in the airways. Guillemette Marot is involved in this H2020 project as scientific head of bilille platform, and will supervise 1 year engineer on integration of omic data.

H2020 PERF-AI

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

- **Acronym:** PERF-AI
- **Project title:** Enhance Aircraft Performance and Optimisation through utilisation of Artificial Intelligence
- **Coordinator:** Pierre Jouniaux (Safety-Line)

- **Duration:** 2 years (2018–2020)
- **Partners:** Safety-Line
- **Abstract:** PERF-AI will apply Machine Learning techniques on flight data (parametric & non-parametric approaches) to accurately measure actual aircraft performance throughout its lifecycle.

  Within current airline operations, both at flight preparation (on-ground) & at flight management (in-air) levels, the trajectory is first planned, then managed by the Flight Management System (FMS) using a single manufacturer's performance model that is the same for every aircraft of the same type, & also on weather forecast that is computed long before the flight. It induces a lack of accuracy during the planning phase with a flight route pre-established at specific altitudes & speeds to optimize fuel burn, from take-off to landing using aircraft performances that are not those of the real aircraft. Also, the actual flight will usually shift from the original plan because of Air Traffic Control (ATC) constraints, adverse weather, wind changes & tactical re-routing, without possibility for the flight crew, either using the FMS or through connected services to tactically recompute the trajectory in order to continuously optimize the flight path. This is in particular due to the limitations of the performance databases that the current systems are using.

  Hence, PERF-AI is focusing on identifying adequate machine learning algorithms, testing their accuracy & capability to perform flight data statistical analysis & developing mathematical models to optimize real flight trajectories with respect to the actual aircraft performance, thus, minimizing fuel consumption throughout the flight.

  The consortium consists of Safety-Line & Inria, having full expertise at Aircraft Performance & Data Science, hence, able to fully propose, test & validate different statistical models that will allow to accurately solve some optimization challenges & implement them in an operational environment.

  PERF-AI total grant request to the CSJU is 568 550 € with total project duration of 24 months.

9.4 National initiatives

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

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

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, Inserm, CHRU Lille)
- **Duration:** 5 years
- **Partners:** FHU Integra and Sanofi

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

CNRS PEPS Blanc — BayesRealForRNN project

Participants Pascal Germain, Vera Shalaeva.

- **Acronym:** BayesRealForRNN
- **Project title:** PAC-Bayesian theory for recurrent neural networks: a control theoretic approach
- **Coordinator:** Mihaly Petreczky (CNRS, UMR 9189 CRIStAL, Université de Lille)
- **Year:** 2019
- **Abstract:** The project proposes to analyze the mathematical correctness of deep learning algorithms by combining techniques from control theory and PAC-Bayesian statistical theory. More precisely, the project proposes to concentrate on recurrent neural networks (RNNs), develop their structure theory using techniques from contol theory, and then apply this structure theory to derive PAC-Bayesian error bounds for RNNs.

CNRS AMIES PEPS 2 — DiagChange project

Participants Cristian Preda *(coordinator)*, Quentin Grimonprez.

- **Acronym:** DiagChange
- **Year:** 2019
- **Abstract:** The project proposes to study the topic of change detection distribution for multivariate signal in a industrial context. The project is in collaboration with the DiagRAMS start-up.

CNRS AMIES PEPS 1 — PIVISCoT

Participants Sophie Dabo-Niang *(coordinator)*, Cristian Preda.

- **Year:** 2020
- **Abstract:** The project aims to create a software for Territorial Coherence Scheme (SCoT) in Lille in order to monitor urban developments and develop territorial observation.

AMIES PEPS 2 — MadiPa

Participants Stéphane Girard, Serge Iovleff *(coordinator)*.

- **Acronym:** MadiPa
- **Project title:** Modèles Auto-associatifs pour la Dispersion de Polluants dans l'Atmosphère
- **Duration:** 18 month (start in december 2019)
- **Partners:** Société Phimeca http://phimeca.com/, Mistis team Inria Grenoble Rhône-Alpes
- **Abstract:** Our goal is to develop a method for predicting the dispersion of pollutants in the atmosphere from an initial emission map and meteorological data. A map of the probabilities of exceeding a critical threshold of pollutants will be estimated thanks to the construction of a meta-model: the large dimension of the problem is reduced by the use of auto-associative models, a non-linear extension of the Principal Components Analysis.

9.4.1 ANR

APRIORI

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

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

BEAGLE

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

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

SMILE

Participants Christophe Biernacki, Vincent Vandewalle.

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

TheraSCUD2022

Participants Guillemette Marot.

- **Acronym:** TheraSCUD2022
- **Project title:** Targeting the IL-20/IL-22 balance to restore pulmonary, intestinal and metabolic homeostasis after cigarette smoking and unhealthy diet
- **Coordinator:** P. Gosset (Institut Pasteur de Lille)
- **Duration:** 3 years (2017–2020)
- **Partners:** CIIL Institut Pasteur de Lille and UMR 1019 INRA Clermont-Ferrand
- **Abstract:** The TheraSCUD2022 project studies inflammatory disorders associated with cigarette smoking and unhealthy diet (SCUD). Guillemette Marot is involved in this ANR project as head of bililleplatform, and will supervise 1 year engineer on integration of omic data.

9.4.2 Working groups

- Sophie Dabo-Niang belongs to the following working groups:
  - STAFAV (STatistiques pour l'Afrique Francophone et Applications au Vivant)
  - ERCIM Working Group on computational and Methodological Statistics, Nonparametric Statistics Team
  - Franco-African IRN (International Research Network) in Mathematics, funded by CNRS
  - ONCOLille (Cancer Research Institute in Lille)
- Benjamin Guedj belongs to the following working groups (GdR) of CNRS:
  - ISIS (local referee for Inria Lille - Nord Europe)
  - MaDICS
  - MASCOT-NUM (local referee for Inria Lille - Nord Europe)
- Guillemette Marot belongs to the StatOmique working group

9.5 Regional initiatives

9.5.1 bilille, the bioinformatics platform of Lille

Participants Guillemette Marot, Maxime Brunin, Iheb Eladib.

bilille, the bioinformatics platform of Lille officially integrated UMS 2014/US 41 PLBS (Plateformes Lilloises en Biologie Santé) in January 2020. In 2020, Guillemette Marot co-headed the platform with Hélène Touzet (CNRS, CRIStAL). Inria employed 2 engineers for this platform:

- M. Brunin, who participated in the development of the visCorVar tool, a tool to facilitate multi-block analysis for statistical integration of omics data and participated to the analyses of the TheraSCUD2022 ANR project.
- I. Eladib, who participated in the development of tools for bilille cloud, in order to simplify and optimize its use.

More information about the platform is available at https://wikis.univ-lille.fr/bilille/

Collaborations of the year linked to bilille

Participants Guillemette Marot.

Guillemette Marot has supervised the data analysis part or support in biostatistics tools testing for the following research projects involving engineers from bilille (only the names of the principal investigators of the project are given even if several partners are sometimes involved in the project):

- CIIL, L. Poulin, InflammReg
- Infinite, V. Sobanski, Evapass
- U1011, Y. Sebti, Circaregen
- U1011, D. Dombrowicz, DeconImmunMetab

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- Benjamin Guedj has been appointed (March 2020) general local chair of COLT 2022 to be held in London
- Hemant Tyagi is the organizer of the MODAL team scientific seminar
- Sophie Dabo-Niang is co-chair of the group *Statistics, applied math and computer science* of *Pan-African Scientific Research Council*, funded by Princeton University (USA)

Member of the organizing committees

- Sophie Dabo-Niang is co-chair of the Organizing Committee of the Workshop *3rd Conference on Econometrics for Environment*, December 2020, Lille.

10.1.2 Scientific events: selection

Christophe Biernacki has been president of the scientific comitee of JdS 2020, the annual national meeting the French staticial society (SFdS).

Reviewer

- Sophie Dabo-Niang has reviewed several papers for several journals during 2020 including Spatial Statistics, JSPI, Metrika, JRSS C
- Benjamin Guedj has served as reviewer for most top-tier machine learning conferences, including AISTATS, ALT, COLT, ICML, NeurIPS
- Hemant Tyagi has reviewed for the following conferences during 2020: International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML) and Symposium on Computational Geometry (SoCG)
- Christophe Biernacki has reviewed for the Cap2020 (Conférence sur l'Apprentissage Automatique) and also for several journals (IMAIAI, STCO, LSSP, SAM, GSCS, TNNLS, ESWA, JMIV, JCGS)

10.1.3 Journal

Member of the editorial boards

- Sophie Dabo-Niang is member of the editorial board of: *Revista Colombiana de Estadística Journal Of Statistical Modeling and Analytics*
- Benjamin Guedj is a member of the Editorial Board of reviewers for the Journal of Machine Learning Research (JMLR), since June 2020 and an Associate Editor and member of the Editorial Board for the journal Information and Inference (Oxford), since March 2020
- Christophe Biernacki is an Associate Editor of the North-Western European Journal of Mathematics (NWEJM) and a Guest Editor for the Special Issue on Innovations in Model-Based Clustering and Classification of the journal Advances Data Analysis and Classification (ADAC)
- Cristian Preda is an Associate Editor for Methodology and Computing in Applied Probability Journal (https://www.springer.com/journal/11009) and Romanian Journal of Mathematics and Computer Science (http://www.rjm-cs.ro)

Reviewing activities

- Hemant Tyagi has reviewed for the following journals during 2020: Journal of the Royal Statistical Society (JRSS), IEEE Open Journal of Signal Processing, Mathematical reviews.
- Vincent Vandewalle has reviewed for the following journals during 2020: JCGS, Spatial Statistics, Methodology & Computing in Applied Probability.

10.1.4 Invited talks

Benjamin Guedj has given a number of scientific talks in seminars, including at

- Oxford University (United Kingdom)
- UCL (United Kingdom)
- The Alan Turing Institute (United Kingdom)
- RIKEN (Japan)

Sophie Dabo-Niang bas been invited to:

- NEF (Next Enstein Forum) 2020, December 8-10, 2020. Panel on *The contribution of Mathematical Sciences in supporting robust disease prevention and modelling in Africa.*
- AIMS South-Africa webinar, November 4, 2020. *Statistical modeling of Spatial Big data and Applications.*

Hemant Tyagi:

- Cafe de Sciences, Inria Lille, January 2020.
- STADIUS seminar, KU Leuven, February 2020.
- Séminaire SAMM : Statistique, Analyse et Modélisation Multidisciplinaire, Université Paris 1, November 2020.

10.1.5 Leadership within the scientific community

Sophie Dabo-Niang is:

- Chair of Committee for Developing Countries (CDC) of EMS (European Mathematical Society), 2019-2022. CDC
- Member of the executif committee and scientif officer of CIMPA

Guillemette Marot is scientific head of bilille, the bioinformatics platform of Lille. More information about the platform is available at `https://wikis.univ-lille.fr/bilille/`

10.1.6 Scientific expertise

Sophie Dabo-Niang is expert of

- L'Oreal *Women in Science* Awards
- HCERES

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Pascal Germain taught
  - Master: Introduction aux réseaux de neurones, 15 heures, M2, Université de Lille, France
- Hemant Tyagi is teaching
  - Master: Statistics I, 24h, M1, Centrale Lille, France (Nov. 2020 - 7 Jan. 2021)
  - Master: Statistics II, 24h, M1, Centrale Lille, France (11 Jan. 2021 - 18 March 2021)
- Sophie Dabo-Niang is teaching
  - Master: Spatial Statistics, 24h, M2, Université de Lille, France
  - Master: Advanced Statistics, 24h, M2, Université de Lille, France
  - Master: Multivariate Data Analyses, 24h, M2, Université de Lille, France
  - Licence: Probability, 24h, L2, Université de Lille, France
  - Licence: Multivariate Statistics, 24h, L3, Université de Lille, France
- Guillemette Marot is teaching
  - Licence: Biostatistics, 15h, L1, Université de Lille (Faculty of Medicine), France
  - Master: Biostatistics, 62h, M1, Université de Lille (Faculty of Medicine), France

  - Master: Supervised classification, 34h, M1, Polytech'Lille, France
  - Master: Biostatistics, 20h, M1, Université de Lille (Departments of Computer Science and Biology), France
  - Master: Statistical analysis of omics data, 22h, M2, Université de Lille (Department of Mathematics), France
  - Doctorat: Artificial intelligence and health, 7h, Université de Lille (Faculty of Medicine), France

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

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

- Benjamin Guedj is teaching
  - Advanced machine learning (M2, 6h), University College London, United Kingdom

- Serge Iovleff is teaching
  - Licence: Analyse et méthodes numériques, 56h, Université de Lille, DUT Informatique
  - Licence: R.O. et aide à la décision, 32h, Université de Lille, DUT Informatique

- Vincent Vandewalle is teaching
  - Licence: Probability, 60h, Université de Lille, DUT STID
  - Licence: Case study in statistics, 45h, Université de Lille, DUT STID
  - Licence: R programming, 45h, Université de Lille, DUT STID
  - Licence: Supervised clustering, 32h, Université de Lille, DUT STID
  - Licence: Analysis, 24h, Université de Lille, DUT STID

10.2.2 Supervision

PhD defense:

- Arthur Leroy, December 9th 2020 on "Apprentissage de donneées fonctionnelles par modèles multi-tâches : application à la prédiction de performances sportives"

- Yaroslav Averyanov, December 15th 2020, supervised by Alain Celisse and Cristian Preda on "Designing and analyzing new early stopping rules for saving computational resources"

- Margot Selosse, November 13th 2020, supervised by Christophe Biernacki and Julien Jacques on "Introducing parsimony to analyse complex data with model-based clustering"

PhD in progress:

- Axel Potier, Sale prediction for low turn-over products, November 2020, Christophe Biernacki, Matthieu Marbac, Vincent Vandewalle
- Felix Biggs, Generative models and kernels, University College London (United Kingdom), Sep 2019, Benjamin Guedj
- Antoine Vendeville, Learning on graph to stop the propagation of fake news, University College London (United Kingdom), Sep 2019, Benjamin Guedj
- Luxin Zhang, Domain adaptation from a pre-trained source model – Application to fraud detection in electronic payments, February 2019, Christophe Biernacki, Pascal Germain, Yacine Kessac
- Paul Viallard, Interpreting representation learning through PAC-Bayes theory, September 2019, Amaury Habrard, Emilie Morvant, Pascal Germain
- Dang Khoi Pham, Planning and re-planning of nurses in an oncology department using a multi-objective and interdisciplinary approach, September 2016, Sophie Dabo-Niang
- Solange Doumun, Performance evaluation and contribution to the development of multispectral image analysis strategies for automatic and rapid diagnosis of malaria, December 2018, Sophie Dabo-Niang
- Alaa Ali Ayad, Statistical modeling of large spatial data and its applications in health, September 2018, Sophie Dabo-Niang
- Wilfried Heyse, Prise en compte de la structure temporelle dans l'analyse statistique de données protéomiques à haut débit, October 2019, Christophe Bauters, Guillemette Marot and Vincent Vandewalle
- Margot Selosse, October 2017, Christophe Biernacki and Julien Jacques
- Filippo Antonazzo, October 2019, Christophe Biernacki and Christine Keribin
- Eglantine Karle, November 2020, Hemant Tyagi and Cristian Preda
- Guillaume Braun, January 2020, Christophe Biernacki and Hemant Tyagi
- Rajeev Bopche, September 2020, Christophe Biernacki and Martine Vaxillaire
- Antonin Schrab, September 2020, co-supervised by Arthur Gretton and Benjamin Guedj, University College London (United Kingdom)
- Reuben Adams, Septembre 2020, co-supervised by John Shawe-Taylor and Benjamin Guedj, University College London (United Kingdom)

10.2.3 Juries

- Sophie Dabo-Niang acted as a reviewer and an examinator for PhD theses
- Benjamin Guedj has been the discussion leader for the licentiate thesis of Fredrik Hellström on December 16th, 2020, at Chalmers University (Sweden)
- Benjamin Guedj has been a member of 2 hiring panels for Inria permanent researchers
- Guillemette Marot acted as an examinator for the PhD thesis of Audrey Hulot, Nov 2020 (Université Paris-Saclay) and in a research engineer (IR) jury, Oct 2020 (Université de Lille)
- Christophe Biernacki acted as a reviewer for four PhD theses and as an examinator for two HdR defenses
- Vincent Vandewalle participated in a MC jury Université d'Avignon, May 2020

- Cristian Preda acted as a referee for the HDR defense of Christophe Crambes, Université de Montpellier 2, June 30, 2020
- Cristian Preda acted as a referee for the HDR defense of Dan Lascu, November 19, 2020, Universitatea Ovidiu, Constanta (Romania)

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 and A. Dermoune. 'Parametrizations, fixed and random effects'. In: *Journal of Multivariate Analysis* 154 (Feb. 2017), pp. 162–176. DOI: 10.1016/j.jmva.2016.11.001. URL: https://hal.archives-ouvertes.fr/hal-01655461.

[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] A. Adekpedjou and S. Dabo-Niang. 'Semiparametric estimation with spatially correlated recurrent events'. In: *Scandinavian Journal of Statistics* (28th June 2020). DOI: 10.1111/sjos.12480. URL: https://hal.inria.fr/hal-03133810.

[12] M.-S. Ahmed, S. Dabo-Niang, M. Genin and A. A. Hassan. 'Partially Linear Spatial Probit Models'. In: *Annales de l'ISUP* (31st Dec. 2020). URL: https://hal.inria.fr/hal-03133818.

[13] H. Alawieh, N. Wicker and C. Biernacki. 'Projection under pairwise distance controls'. In: *Communications in Statistics - Theory and Methods* (2020). DOI: 10.1080/03610926.2020.1741626. URL: https://hal.archives-ouvertes.fr/hal-01420662.

[14] M. A. B. Alaya, C. Ternynck, S. Dabo-Niang, F. Chebana and T. B. Ouarda. 'Change point detection of flood events using a functional data framework'. In: *Advances in Water Resources* 137 (Mar. 2020), p. 103522. DOI: 10.1016/j.advwatres.2020.103522. URL: https://hal.inria.fr/hal-03133809.

[15] C. Biernacki, M. Marbac and V. Vandewalle. 'Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering'. In: *Journal of Classification* (11th July 2020). DOI: 10.1007/s00357-020-09369-y. URL: https://hal.archives-ouvertes.fr/hal-01949155.

[16] S. Boumeddane, L. Hamdad, H. Haddadou and S. Dabo-Niang. 'A kernel discriminant analysis for spatially dependent data'. In: *Distributed and Parallel Databases* (27th Aug. 2020). DOI: 10.1007/s10619-020-07309-8. URL: https://hal.inria.fr/hal-03133813.

[17] S. Chrétien and H. Tyagi. 'Multi-kernel unmixing and super-resolution using the Modified Matrix Pencil method'. In: *Journal of Fourier Analysis and Applications* 26.18 (22nd Jan. 2020). DOI: 10.1007/s00041-020-09725-x. URL: https://hal.inria.fr/hal-02379598.

[18] M. Cucuringu and H. Tyagi. 'Provably robust estimation of modulo 1 samples of a smooth function with applications to phase unwrapping'. In: *Journal of Machine Learning Research* 21.32 (1st Jan. 2020), 1–77. URL: https://hal.inria.fr/hal-02379573.

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

[20] S. Dabo-Niang and B. Thiam. 'Kernel regression estimation with errors-in-variables for random fields'. In: *Afrika Matematika* 31 (2020), pp. 29–56. DOI: 10.1007/s13370-019-00654-7. URL: https://hal.inria.fr/hal-02334993.

[21] F. Dewez, B. Guedj and V. Vandewalle. 'From industry-wide parameters to aircraft-centric on-flight inference: improving aeronautics performance prediction with machine learning'. In: *Data-Centric Engineering* (19th Oct. 2020). DOI: 10.1017/dce.2020.12. URL: https://hal.inria.fr/hal-02570875.

[22] S. Doumun, S. Dabo-Niang and J. Zoueu. 'Detection and segmentation of erythrocytes in multispectral label-free blood smear images for automatic cell counting'. In: *Journal of Spectral Imaging* 9.Article ID a10 (9th Sept. 2020). DOI: 10.1255/jsi.2020.a10. URL: https://hal.inria.fr/hal-03133812.

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

[24] P. Germain, A. Habrard, F. Laviolette and E. Morvant. 'PAC-Bayes and Domain Adaptation'. In: *Neurocomputing* 379 (2020), pp. 379–397. DOI: 10.1016/j.neucom.2019.10.105. URL: https://hal.archives-ouvertes.fr/hal-01563152.

[25] B. Guedj and B. S. Desikan. 'Kernel-Based Ensemble Learning in Python'. In: *Information* 11.2 (Feb. 2020), p. 63. DOI: 10.3390/info11020063. URL: https://hal.inria.fr/hal-02443097.

[26] C. Preda and A. Amarioarei. 'One Dimensional Discrete Scan Statistics for Dependent Models and Some Related Problems'. In: *Mathematics* 8.4 (Apr. 2020), p. 576. DOI: 10.3390/math8040576. URL: https://hal.archives-ouvertes.fr/hal-03114193.

[27] M. Selosse, J. Jacques and C. Biernacki. 'Model-based co-clustering for mixed type data'. In: *Computational Statistics and Data Analysis* 144 (2020), p. 106866. DOI: 10.1016/j.csda.2019.106866. URL: https://hal.archives-ouvertes.fr/hal-01893457.

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

[29] M. Selosse, J. Jacques and C. Biernacki. 'Textual data summarization using the Self-Organized Co-Clustering model'. In: *Pattern Recognition* (Feb. 2020). DOI: 10.1016/j.patcog.2020.107315. URL: https://hal.archives-ouvertes.fr/hal-02115294.

[30] D. Souami, J. Cresson, C. Biernacki and F. Pierret. 'On the local and global properties of the gravitational spheres of influence'. In: *Monthly Notices of the Royal Astronomical Society* 496.4 (2nd June 2020), pp. 4287–429. DOI: 10.1093/mnras/staa1520. URL: https://hal.inria.fr/hal-02617073.

[31] V. Vandewalle. 'Multi-Partitions Subspace Clustering'. In: *Mathematics* 8.4 (Apr. 2020), p. 597. DOI: 10.3390/math8040597. URL: https://hal.inria.fr/hal-03117603.

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

International peer-reviewed conferences

[33] F. Antonazzo, C. Biernacki and C. Keribin. 'A binned technique for scalable model-based clustering on huge datasets'. In: MBC2 - Models and Learning for Clustering and Classification. Journal ADAC - Advances in Data Analysis and Classification, Catania, Italy, 2nd Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-03097284.

[34] G. Braun, C. Biernacki and H. Tyagi. 'Clustering on multilayer graphs with missing values'. In: Journée de Statistique de la SFdS. Nice, France, 25th May 2020. URL: https://hal.inria.fr/hal-03122104.

[35] S. Chretien and B. Guedj. 'Revisiting clustering as matrix factorisation on the Stiefel manifold'. In: LOD 2020 - the Sixth International Conference on Machine Learning, Optimisation and Data Science. Siena, Italy, 19th July 2020. URL: https://hal.inria.fr/hal-02064396.

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

[37] B. Guedj and J. Rengot. 'Non-linear aggregation of filters to improve image denoising'. In: Computing Conference 2020. London, United Kingdom, 16th July 2020. URL: https://hal.inria.fr/hal-02086856.

[38] W. Heyse, V. Vandewalle, P. Amouyel, G. Marot, C. Bauters and F. Pinet. 'Proteomic signature for early diagnosis of left ventricular remodeling after myocardial infarction'. In: Printemps de la cardiologie 2020. Grenoble, France, 29th Oct. 2020. URL: https://hal.inria.fr/hal-03124801.

[39] Z. Mhammedi, B. Guedj and R. C. Williamson. 'PAC-Bayesian Bound for the Conditional Value at Risk'. In: NeurIPS 2020. Vancouver / Virtual, Canada, 6th Dec. 2020. URL: https://hal.inria.fr/hal-02883728.

[40] K. Nozawa, P. Germain and B. Guedj. 'PAC-Bayesian Contrastive Unsupervised Representation Learning'. In: UAI 2020 - Conference on Uncertainty in Artificial Intelligence. Toronto, Canada, 3rd Aug. 2020. URL: https://hal.inria.fr/hal-02401282.

[41] V. Shalaeva, A. Fakhrizadeh Esfahani, P. Germain and M. Petreczky. 'Improved PAC-Bayesian Bounds for Linear Regression'. In: AAAI 2020 - Thirty-Fourth AAAI Conference on Artificial Intelligence. New York, United States, 7th Feb. 2020. URL: https://hal.inria.fr/hal-02396556.

[42] L. Zhang, P. Germain, Y. Kessaci and C. Biernacki. 'Target to Source Coordinate-wise Adaptation of Pre-trained Models'. In: ECML PKDD 2020 - The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Ghent / Virtual, Belgium, 14th Sept. 2020. URL: https://hal.inria.fr/hal-03087284.

National peer-reviewed Conferences

[43] F. Antonazzo, C. Biernacki and C. Keribin. 'Estimation of univariate Gaussian mixtures for huge raw datasets by using binned datasets'. In: JDS 2020 - 52ème Journées de Statistiques de la Société Française de Statistique. Nice, France, 25th May 2020. URL: https://hal.archives-ouvertes.fr/hal-03082437.

Conferences without proceedings

[44] A. Amarioarei and C. Preda. 'Scan statistics for some dependent models.Applications.' In: STATMOD2020 Statistical Modeling with Applications. Bucharest, Romania, 15th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03141585.

[45] S. Dabo-Niang. 'The contribution of Mathematical Sciences in supporting robust disease prevention and modelling in Africa'. In: The contribution of Mathematical Sciences in supporting robust disease prevention and modelling in Africa. Virtual Meeting, South Africa, 10th Dec. 2020. URL: https://hal.inria.fr/hal-03133823.

[46] M. Selosse, I. C. Gormley, J. Jacques and C. Biernacki. 'A bumpy journey: exploring deep Gaussian mixture models'. In: I Can't Believe It's Not Better @ NeurIPS 2020. Vancouver, Canada, 12th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-02985701.

[47] M. Selosse, J. Jacques and C. Biernacki. 'Co-clustering contraint pour le résumé de matrices document-terme'. In: JdS 2020 - 52èmes Journées de Statistique de la Société Française de Statistique. Nice, France, 25th May 2020. URL: https://hal.archives-ouvertes.fr/hal-02490028.

[48] V. Vandewalle, A. Caron and B. Dervaux. 'Estimation du nombre de problèmes et détermination du nombre de sujets nécessaires dans les études d'utilisabilité : une approche bayésienne'. In: Journées Biostatistiques 2020 - GDR « Statistiques & santé ». Paris, France, 1st Oct. 2020. URL: https://hal.inria.fr/hal-03118164.

Scientific book chapters

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

Edition (books, proceedings, special issue of a journal)

[50] S. Dabo-Niang, S. Manou-Abi and S. Jean-Jacques. *Mathematical Modeling and Study of Random or Deterministic Phenomena*. Wiley, 1st Feb. 2020. URL: https://hal.inria.fr/hal-02334997.

Doctoral dissertations and habilitation theses

[51] Y. Averyanov. 'Designing and analyzing new early stopping rules for saving computational resources'. Université de Lille; Inria, 15th Dec. 2020. URL: https://tel.archives-ouvertes.fr/tel-03133391.

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

Reports & preprints

[53] F. Biggs and B. Guedj. *Differentiable PAC-Bayes Objectives with Partially Aggregated Neural Networks*. 23rd June 2020. URL: https://hal.inria.fr/hal-02879216.

[54] L. Brotcorne, A. Canteaut, A. C. Viana, C. Grandmont, B. Guedj, S. Huot, V. Issarny, G. Pallez, V. Perrier, V. Quema, J.-B. Pomet, X. Rival, S. Salvati and E. Thomé. *Indicateurs de suivi de l'activité scientifique de l'Inria*. Inria, 1st Dec. 2020. URL: https://hal.inria.fr/hal-03033764.

[55] T. Cantelobre, B. Guedj, M. Pérez-Ortiz and J. Shawe-Taylor. *A PAC-Bayesian Perspective on Structured Prediction with Implicit Loss Embeddings*. 8th Dec. 2020. URL: https://hal.inria.fr/hal-03046401.

[56] A. Celisse and M. Wahl. *Analyzing the discrepancy principle for kernelized spectral filter learning algorithms*. 10th Apr. 2020. URL: https://hal.inria.fr/hal-02548917.

[57] M. Cucuringu, A. V. Singh, D. Sulem and H. Tyagi. *Regularized spectral methods for clustering signed networks*. 7th Jan. 2021. URL: https://hal.inria.fr/hal-03101710.

[58] M. Cucuringu and H. Tyagi. *An extension of the angular synchronization problem to the heterogeneous setting*. 7th Jan. 2021. URL: https://hal.inria.fr/hal-03101682.

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

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

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

[62] F. Dewez, B. Guedj, A. Talpaert and V. Vandewalle. *An end-to-end data-driven optimisation framework for constrained trajectories*. 25th Nov. 2020. URL: https://hal.inria.fr/hal-03024720.

[63] A. Ehrhardt, C. Biernacki, V. Vandewalle, P. Heinrich and S. Beben. *Reject Inference Methods in Credit Scoring: A rational review*. 23rd Dec. 2020. URL: https://hal.inria.fr/hal-03087279.

[64] M. Fanuel and H. Tyagi. *Denoising modulo samples: k-NN regression and tightness of SDP relaxation*. 7th Jan. 2021. URL: https://hal.inria.fr/hal-03101740.

[65] M. P. B. Gallaugher, C. Biernacki and P. D. McNicholas. *Parameter-Wise Co-Clustering for High-Dimensional Data*. 30th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-01862824.

[66] M. Haddouche, B. Guedj, O. Rivasplata and J. Shawe-Taylor. *PAC-Bayes unleashed: generalisation bounds with unbounded losses*. 17th June 2020. URL: https://hal.inria.fr/hal-02872173.

[67] M. Haddouche, B. Guedj, O. Rivasplata and J. Shawe-Taylor. *Upper and Lower Bounds on the Performance of Kernel PCA*. 21st Dec. 2020. URL: https://hal.inria.fr/hal-03084598.

[68] S. Iovleff, S. N. Sylla and C. Loucoubar. *Block clustering of Binary Data with Gaussian Co-variables*. 1st Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-01961978.

[69] A. Leroy, P. Latouche, B. Guedj and S. Gey. *Cluster-Specific Predictions with Multi-Task Gaussian Processes*. 17th Nov. 2020. URL: https://hal.inria.fr/hal-03009276.

[70] A. Leroy, P. Latouche, B. Guedj and S. Gey. *MAGMA: Inference and Prediction with Multi-Task Gaussian Processes*. 22nd July 2020. URL: https://hal.inria.fr/hal-02904446.

[71] M. Marbac, M. Sedki, C. Biernacki and V. Vandewalle. *Simultaneous semi-parametric estimation of clustering and regression*. 29th Dec. 2020. URL: https://hal.inria.fr/hal-03090573.

[72] C. Preda, Q. Grimonprez and V. Vandewalle. *cfda: an R Package for Categorical Functional Data Analysis*. 20th Oct. 2020. URL: https://hal.inria.fr/hal-02973094.

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

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

[75] H. Tyagi. *Error analysis for denoising smooth modulo signals on a graph*. 7th Jan. 2021. URL: https://hal.inria.fr/hal-03101720.

[76] A. Vendeville, B. Guedj and S. Zhou. *Forecasting elections results via the voter model with stubborn nodes*. 23rd Sept. 2020. URL: https://hal.inria.fr/hal-02946434.

[77] A. Vendeville, B. Guedj and S. Zhou. *How opinions crystallise: an analysis of polarisation in the voter model*. 17th June 2020. URL: https://hal.inria.fr/hal-02872161.

Other scientific publications

[78] C. Biernacki and V. Vandewalle. *Label switching in mixtures*. Glasgow, United Kingdom, France, 17th July 2021. URL: https://hal.inria.fr/hal-03183299.

[79] W. Heyse, V. Vandewalle, P. Amouyel, G. Marot, C. Bauters and F. Pinet. *Proteomic signature for early diagnosis of left ventricular remodeling after myocardial infarction*. Grenoble / Virtual, France, 29th Oct. 2020. URL: https://hal.inria.fr/hal-03124837.

11.3 Cited publications

[80] V. Barbu and N. Limnios. 'Reliability theory for discrete-time semi-Markov systems'. In: *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*. Springer, 2008, pp. 1–30.

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

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

[83] Z. Harchaoui and O. Cappé. 'Retrospective Mutiple Change-Point Estimation with Kernels'. In: *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*. 2007, pp. 768–772. DOI: 10.1109/SSP.2007.4301363.

[84] K. Kawaguchi, L. P. Kaelbling and Y. Bengio. 'Generalization in Deep Learning'. In: *CoRR* abs/1710.05468 (2017). arXiv: 1710.05468. URL: http://arxiv.org/abs/1710.05468.

[85] E. Lebarbier. 'Detecting multiple change-points in the mean of Gaussian process by model selection'. In: *Signal Processing* 85.4 (2005), pp. 717–736. DOI: https://doi.org/10.1016/j.sigpro.2004.11.012. URL: https://www.sciencedirect.com/science/article/pii/S0165168404003196.

[86] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak and H. Khandeparkar. 'A Theoretical Analysis of Contrastive Unsupervised Representation Learning'. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 5628–5637. URL: http://proceedings.mlr.press/v97/saunshi19a.html.

[87] S. Trevezas and N. Limnios. 'Exact MLE and asymptotic properties for nonparametric semi-Markov models'. In: *Journal of Nonparametric Statistics* 23.3 (2011), pp. 719–739.