RESEARCH CENTRE

Sophia Antipolis - Méditerranée

IN PARTNERSHIP WITH:

Université Côte d'Azur

2021 ACTIVITY REPORT

Project-Team MAASAI

Models and Algorithms for Artificial Intelligence

IN COLLABORATION WITH: Laboratoire informatique, signaux systèmes de Sophia Antipolis (I3S), Laboratoire Jean-Alexandre Dieudonné (JAD)

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Contents

Project-Team MAASAI

Creation of the Project-Team: 2020 February 01

Keywords

Computer sciences and digital sciences

A3.1. – Data

A3.1.10. – Heterogeneous data

A3.1.11. – Structured data

A3.4. – Machine learning and statistics

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.6. – Neural networks

A3.4.7. – Kernel methods

A3.4.8. – Deep learning

A9. – Artificial intelligence

A9.2. – Machine learning

Other research topics and application domains

B6.3.4. – Social Networks

B7.2.1. – Smart vehicles

B8.2. – Connected city

B9.6. – Humanities

1 Team members, visitors, external collaborators

Research Scientist

- Pierre Alexandre Mattei [Inria, Researcher]

Faculty Members

- Charles Bouveyron [Team leader, Université Côte d'Azur, Professor, HDR]
- Marco Corneli [Université Côte d'Azur, Associate Professor]
- Damien Garreau [Université Côte d'Azur, Associate Professor]
- Marco Gori [Universita di Siena, Italy, Professor]
- Frederic Precioso [Université Côte d'Azur, Professor]
- Michel Riveill [Université Côte d'Azur, Professor]

Post-Doctoral Fellows

- Juliette Chevallier [Université Côte d'Azur, until July 2021]
- Aude Sportisse [Inria, from Oct 2021]
- Gabriel Wallin [Université Côte d'Azur, until July 2021]

PhD Students

- Gabriele Ciravegna [Université de Vérone-Italie, from Jun 2021 until Oct 2021]
- Celia Dcruz [Université Côte d'Azur, from Oct 2021]
- Kevin Dsouza [Université Côte d'Azur, from Oct 2021]
- Dingge Liang [Université Côte d'Azur]
- Gianluigi Lopardo [Université Côte d'Azur, from Oct 2021]
- Giulia Marchello [Université Côte d'Azur]
- Taki Eddine Mekhalfa [Université Côte d'Azur, until May 2021]
- Hugo Miralles [Orange, CIFRE]
- Kevin Mottin [Université Côte d'Azur, from Nov 2021]
- Louis Ohl [Université Côte d'Azur, from Mar 2021]
- Baptiste Pouthier [NXP, Cifre]
- Miguel Romero Rondon [Université Côte d'Azur, until Sep 2021]
- Laura Sanabria Rosas [Université Côte d'Azur]
- Hugo Schmutz [Université Côte d'Azur]
- Cedric Vincent-Cuaz [Université Côte d'Azur]
- Xuchun Zhang [Université Côte d'Azur]
- Mansour Zoubeirou A Mayaki [Pro BTP, CIFRE]

Technical Staff

- Stephane Petiot [Inria, Engineer, from Mar 2021]

Interns and Apprentices

- Julien Choukroun [Univ Côte d'Azur, from Jun 2021 until Sep 2021]
- Oumar Dieng [Univ Côte d'Azur, from Mar 2021 until Sep 2021]
- Audrey Freysse [Université de Bordeaux, Intern, from 01/01/2021 to 31/08/2021, Master de Pharmacologie]
- Gustavo Gavanzo Chaves [Univ Côte d'Azur, from Mar 2021 until Jul 2021]
- Lara Herrmann [Inria, from Feb 2021 until Jul 2021]
- Mohamed Issa [Univ Côte d'Azur, from Jun 2021 until Aug 2021]
- Jeremy Lombard [Inria, from Apr 2021 until Sep 2021]
- Gianluigi Lopardo [Inria, from Mar 2021 until Sep 2021]
- Anthony Michelard [École d'ingénieur Polytech de Nice-Sophia, from Jun 2021 until Sep 2021]
- Louis Ohl [INSA Lyon, Jan 2021]
- Serena Romani [Université de Bordeaux, Intern, from 01/01/2021 to 30/06/2021, Master de Pharmacologie]
- Jasmine Singh [Univ Côte d'Azur, from Apr 2021 until Jul 2021]
- Julie Tores [Univ Côte d'Azur, from Jun 2021 until Sep 2021]
- Li Yang [Univ Côte d'Azur, from Mar 2021 until Sep 2021]
- Tibou Yao [Univ Côte d'Azur, from Apr 2021 until Jul 2021]

Administrative Assistants

- Nathalie Brillouet [Inria, until Mar 2021]
- Claire Senica [Inria, from Apr 2021]

Visiting Scientists

- Isaure Gonzales Rivera De La Vernhe [Univ Côte d'Azur, from Nov 2021]
- Katherine Minker [Univ Côte d'Azur, from Nov 2021]

External Collaborators

- Hans Greger Ottosson [IBM France, from Aug 2021]
- Samuel Vaiter [CNRS, from Sep 2021]

2 Overall objectives

Artificial intelligence has become a key element in most scientific fields and is now part of everyone life thanks to the digital revolution. Statistical, machine and deep learning methods are involved in most scientific applications where a decision has to be made, such as medical diagnosis, autonomous vehicles or text analysis. The recent and highly publicized results of artificial intelligence should not hide the remaining and new problems posed by modern data. Indeed, despite the recent improvements due to deep learning, the nature of modern data has brought new specific issues. For instance, learning with high-dimensional, atypical (networks, functions, ...), dynamic, or heterogeneous data remains difficult for theoretical and algorithmic reasons. The recent establishment of deep learning has also opened new questions such as: How to learn in an unsupervised or weakly-supervised context with deep architectures? How to design a deep architecture for a given situation? How to learn with evolving and corrupted data?

To address these questions, the Maasai team focuses on topics such as unsupervised learning, theory of deep learning, adaptive and robust learning, and learning with high-dimensional or heterogeneous data. The Maasai team conducts a research that links practical problems, that may come from industry or other scientific fields, with the theoretical aspects of Mathematics and Computer Science. In this spirit, the Maasai project-team is totally aligned with the "Core elements of AI" axis of the Institut 3IA Côte d'Azur. It is worth noticing that the team hosts three 3IA chairs of the Institut 3IA Côte d'Azur, as well as several PhD students funded by the Institut.

3 Research program

Within the research strategy explained above, the Maasai project-team aims at developing statistical, machine and deep learning methodologies and algorithms to address the following four axes.

Unsupervised learning The first research axis is about the development of models and algorithms designed for unsupervised learning with modern data. Let us recall that unsupervised learning — the task of learning without annotations — is one of the most challenging learning challenges. Indeed, if supervised learning has seen emerging powerful methods in the last decade, their requirement for huge annotated data sets remains an obstacle for their extension to new domains. In addition, the nature of modern data significantly differs from usual quantitative or categorical data. We ambition in this axis to propose models and methods explicitly designed for unsupervised learning on data such as high-dimensional, functional, dynamic or network data. All these types of data are massively available nowadays in everyday life (omics data, smart cities, ...) and they remain unfortunately difficult to handle efficiently for theoretical and algorithmic reasons. The dynamic nature of the studied phenomena is also a key point in the design of reliable algorithms.

On the one hand, we direct our efforts towards the development of unsupervised learning methods (clustering, dimension reduction) designed for specific data types: high-dimensional, functional, dynamic, text or network data. Indeed, even though those kinds of data are more and more present in every scientific and industrial domains, there is a lack of sound models and algorithms to learn in an unsupervised context from such data. To this end, we have to face problems that are specific to each data type: How to overcome the curse of dimensionality for high-dimensional data? How to handle multivariate functional data / time series? How to handle the activity length of dynamic networks? On the basis of our recent results, we ambition to develop generative models for such situations, allowing the modeling and the unsupervised learning from such modern data.

On the other hand, we focus on deep generative models (statistical models based on neural networks) for clustering and semi-supervised classification. Neural network approaches have demonstrated their efficiency in many supervised learning situations and it is of great interest to be able to use them in unsupervised situations. Unfortunately, the transfer of neural network approaches to the unsupervised context is made difficult by the huge amount of model parameters to fit and the absence of objective quantity to optimize in this case. We therefore study and design model-based deep learning methods that can hande unsupervised or semi-supervised problems in a statistically grounded way.

Finally, we also aim at developing explainable unsupervised models that can ease the interaction with the practitioners and their understanding of the results. There is an important need for such models,

in particular when working with high-dimensional or text data. Indeed, unsupervised methods, such as clustering or dimension reduction, are widely used in application fields such as medicine, biology or digital humanities. In all these contexts, practitioners are in demand of efficient learning methods which can help them to make good decisions while understanding the studied phenomenon. To this end, we aim at proposing generative and deep models that encode parsimonious priors, allowing in turn an improved understanding of the results.

Understanding (deep) learning models The second research axis is more theoretical, and aims at improving our understanding of the behaviour of modern machine learning models (including, but not limited to, deep neural networks). Although deep learning methods and other complex machine learning models are obviously at the heart of artificial intelligence, they clearly suffer from an overall weak knowledge of their behaviour, leading to a general lack of understanding of their properties. These issues are barriers to the wide acceptance of the use of AI in sensitive applications, such as medicine, transportation, or defense. We aim at combining statistical (generative) models with deep learning algorithms to justify existing results, and allow a better understanding of their performances and their limitations.

We particularly focus on researching ways to understand, interpret, and possibly explain the predictions of modern, complex machine learning models. We both aim at studying the empirical and theoretical properties of existing techniques (like the popular LIME), and at developing new frameworks for interpretable machine learning (for example based on deconvolutions or generative models). Among the relevant application domains in this context, we focus notably on text and biological data.

Another question of interest is: what are the statistical properties of deep learning models and algorithms? Our goal is to provide a statistical perspective on the architectures, algorithms, loss functions and heuristics used in deep learning. Such a perspective can reveal potential issues in exisiting deep learning techniques, such as biases or miscalibration. Consequently, we are also interested in developing statistically principled deep learning architectures and algorithms, which can be particularly useful in situations where limited supervision is available, and when accurate modelling of uncertainties is desirable.

Adaptive and Robust Learning The third research axis aims at designing new learning algorithms which can learn incrementally, adapt to new data and/or new context, while providing predictions robust to biases even if the training set is small.

For instance, we have designed an innovative method of so-called cumulative learning, which allows to learn a convolutional representation of data when the learning set is (very) small. The principle is to extend the principle of Transfer Learning, by not only training a model on one domain to transfer it once to another domain (possibly with a fine-tuning phase), but to repeat this process for as many domains as available. We have evaluated our method on mass spectrometry data for cancer detection. The difficulty of acquiring spectra does not allow to produce sufficient volumes of data to benefit from the power of deep learning. Thanks to cumulative learning, small numbers of spectra acquired for different types of cancer, on different organs of different species, all together contribute to the learning of a deep representation that allows to obtain unequalled results from the available data on the detection of the targeted cancers. This extension of the well-known Transfer Learning technique can be applied to any kind of data.

We also investigate active learning techniques. We have for example proposed an active learning method for deep networks based on adversarial attacks. An unlabelled sample which becomes an adversarial example under the smallest perturbations is selected as a good candidate by our active learning strategy. This does not only allow to train incrementally the network but also makes it robust to the attacks chosen for the active learning process.

Finally, we address the problem of biases for deep networks by combining domain adaptation approaches with Out-Of-Distribution detection techniques.

Learning with heterogeneous and corrupted data The last research axis is devoted to making machine learning models more suitable for real-world, "dirty" data. Real-world data rarely consist in a single kind of Euclidean features, and are genereally heterogeneous. Moreover, it is common to find some form of

corruption in rea-world data sets: for example missing values, outliers, label noise, or even adversarial examples.

Heterogeneous and non-Euclidean data are indeed part of the most important and sensitive applications of artificial intelligence. As a concrete example, in medicine, the data recorded on a patient in an hospital range from images to functional data and networks. It is obviously of great interest to be able to account for all data available on the patients to propose a diagnostic and an appropriate treatment. Notice that this also applies to autonomous cars, digital humanities and biology. Proposing unified models for heterogeneous data is an ambitious task, but first attempts (e.g. the Linkage[1] project) on combination of two data types have shown that more general models are feasible and significantly improve the performances. We also address the problem of conciliating structured and non-structured data, as well as data of different levels (individual and contextual data).

On the basis of our previous works (notably on the modeling of networks and texts), we first intend to continue to propose generative models for (at least two) different types of data. Among the target data types for which we would like to propose generative models, we can cite images and biological data, networks and images, images and texts, and texts and ordinal data. To this end, we explore modelings trough common latent spaces or by hybridizing several generative models within a global framework. We are also interested in including potential corruption processes into these heterogeneous generative models. For example, we are developping new models that can handle missing values, under various sorts of missingness assumptions.

Besides the modelling point of view, we are also interested in making existing algorithms and implementations more fit for "dirty data". We study in particular ways to robustify algorithms, or to improve heuristics that handle missing/corrupted values or non-Euclidean features.

4 Application domains

The Maasai research team has the following major application domains:

Medicine Most of team members apply their research work to Medicine or extract theoretical AI problems from medical situations. In particular, our main applications to Medicine are concerned with pharmacovigilance, medical imaging, and omics. It is worth noticing that medical applications cover all research axes of the team due to the high diversity of data types and AI questions. It is therefore a preferential field of application of the models and algorithms developed by the team.

Digital humanities Another important application field for Maasai is the increasingly dynamic one of digital humanities. It is a extremely motivating field due to the very original questions that are addressed. Indeed, linguists, sociologists, geographers and historians have questions that are quite different than the usual ones in AI. This allows the team to formalize original AI problems that can be generalized to other fields, allowing to indirectly contribute to the general theory and methodology of AI.

Multimedia The last main application domain for Maasai is multimedia. With the revolution brought to computer vision field by deep learning techniques, new questions have appeared such as combining subsymbolic and symbolic approaches for complex semantic and perception problems, or as edge AI to embed machine learning approaches for multimedia solutions preserving privacy. This domain brings new AI problems which require to bridge the gap between different views of AI.

Other application domains Other topics of interest of the team include astronomy, bioinformatics, recommender systems and ecology.

[1] The Linkage project: linkage.fr

5 Highlights of the year

5.1 Funding

- Pierre-Alexandre Mattei was granted a 3IA chair from Institut 3IA Côte d'Azur.
- Greger Ottosson, from IBM, was granted an associate chair from Institut 3IA Côte d'Azur to work 20% of his time within the Maasai team.
- Damien Garreau was granted an ANR JCJC project on the topic of interpretability of machine learning models.

5.2 Awards

- Pierre-Alexandre Mattei received a reviewer award from the International Conference on Machine Learning (ranked among the top 10% of reviewers).

5.3 Conferences co-organised by team members

- Pierre-Alexandre Mattei co-organised the second edition of the Generative Models and Uncertainty Quantification (GenU) workshop. This small-scale workshop was held physically in Copenhagen (October 12-13).
- The team has been deeply involved in the organisation of the 52èmes Journées de la Statistique of the French Statistical Society which have been held (virtualy) in Nice in June 2021. Charles Bouveyron was the President of the organisation committee.
- Charles Bouveyron and Frédéric Precioso were members of the scientific committee of the SophIA-Summit conference.
- Frédéric Precioso is the funder and main organizer of the Deep-Learning School, which gather in July 2021 more than 300 participants online.

5.4 Innovation and transfer

- The Videtics startup has been selected to be supported by the Startitup program of Institut 3IA Côte d'Azur to valorize and transfer knowledges and technologies from the team. C. Bouveyron and F. Precioso are involved in the project.
- The Instant System company has been selected by Inria to benefit from the National "Plan de Relance" and will work with Frédéric Precioso and the team to develop and transfer new AI tools for mobility.
- A contract has been signed with the company Naval Group. The goal of this project will be the development of an open-source Python library for semi-supervised learning, via the hiring of a research engineer.

5.5 Nominations

- Charles Bouveyron has been nominated Director of the Institut 3IA Côte d'Azur, effective January 1st, 2021.

6 New software and platforms

For the Maasai research team, the main objective of the software implementations is to experimentally validate the results obtained and ease the transfer of the developed methodologies to industry. Most of the software will be released as R or Python packages that requires only a light maintaining, allowing a relative longevity of the codes. Some platforms are also proposed to ease the use of the developed methodologies by users without a strong background in Machine Learning, such as scientists from other fields.

The team maintains several R and Python packages, among which the following ones have been released or updated in 2021:

- DAMDA: github.com/michaelfop/damda/
- Fisher-EM: cran.r-project.org/web/packages/FisherEM/index.html
- FunLBM: cran.r-project.org/web/packages/funLBM/index.html
- FunFEM: cran.r-project.org/web/packages/funFEM/index.html
- FunHDDC: cran.r-project.org/web/packages/funHDDC/index.html

The team is also leading the Linkage platform (linkage.fr) which implements an AI technology devoted to the analysis and monitoring of communication networks. The Linkage project is in particular supported by both INRIA, through Plan "National IA", and the Ministry of Research and Higher Education, which provide two full-time engineers.

7 New results

7.1 Unsupervised learning

7.1.1 Latent Class Analysis: Insights about design and analysis of schistosomiasis diagnostic studies

Participants: Elena Erosheva.

Keywords: latent class analysis, schistosomiasis, epidemiology
Collaborations: Artemis Koukounari (F. Hoffmann-La Roche Ltd., United Kingdom), Haziq Jamil (Universiti Brunei Darussalam, Brunei), Clive Shiff (John Hopkins Bloomberg School of Public Health), Irini Moustaki (London School of Economics and Political Science)

Various global health initiatives are currently advocating the elimination of schistosomiasis within the next decade. Schistosomiasis is a highly debilitating tropical infectious disease with severe burden of morbidity and thus operational research accurately evaluating diagnostics that quantify the epidemic status for guiding effective strategies is essential. Latent class models (LCMs) have been generally considered in epidemiology and in particular in recent schistosomiasis diagnostic studies as a flexible tool for evaluating diagnostics because assessing the true infection status (via a gold standard) is not possible. However, within the biostatistics literature, classical LCM have already been criticised for real-life problems under violation of the conditional independence (CI) assumption and when applied to a small number of diagnostics (i.e. most often 3-5 diagnostic tests). Solutions of relaxing the CI assumption and accounting for zero-inflation, as well as collecting partial gold standard information, have been proposed, offering the potential for more robust model estimates. In the [21], we examined such approaches in the context of schistosomiasis via analysis of two real datasets and extensive simulation studies. Our main conclusions highlighted poor model fit in low prevalence settings and the necessity of collecting partial gold standard information in such settings in order to improve the accuracy and reduce bias of sensitivity and specificity estimates.

7.1.2 A New Class of EM Algorithms. Escaping Local Maxima and Handling Intractable Sampling

Participants: Juliette Chevallier.

Keywords: mixture models, variational inference
Collaborations: V. Debavelaere (CMAP), Stéphanie Allassonnière (Univ. de Paris)

The expectation-maximization (EM) algorithm is a powerful computational technique for maximum likelihood estimation in incomplete data models. When the expectation step cannot be performed in closed form, a stochastic approximation of EM (SAEM) can be used. The convergence of the SAEM toward critical points of the observed likelihood has been proved and its numerical efficiency has been demonstrated. However, sampling from the posterior distribution may be intractable or have a high computational cost. Moreover, despite appealing features, the limit position of this algorithm can strongly depend on its starting one. To cope with this two issues, we propose in [11] new stochastic approximation version of the EM in which we do not sample from the exact distribution in the expectation phase of the procedure. We first prove the convergence of this algorithm toward critical points of the observed likelihood. Then, we propose an instantiation of this general procedure to favor convergence toward global maxima. Experiments on synthetic and real data highlight the performance of this algorithm in comparison to the SAEM and the EM when feasible.

7.1.3 Co-Clustering of Multivariate Functional Data for Air Pollution Analysis

Participants: Charles Bouveyron.

Keywords: generative models, model-based co-clustering, functional data, air pollution, public health
Collaborations: J. Jacques and A. Schmutz (Univ. de Lyon), Fanny Simoes and Silvia Bottini (MDlab, MSI, Univ. Côte d'Azur)

In [14], we focused on Air pollution, which is nowadays a major treat for public health, with clear links with many diseases, especially cardiovascular ones. The spatio-temporal study of pollution is of great interest for governments and local authorities when deciding for public alerts or new city policies against pollution raise. The aim of this work is to study spatio-temporal profiles of environmental data collected in the south of France (Région Sud) by the public agency AtmoSud. The idea is to better understand the exposition to pollutants of inhabitants on a large territory with important differences in term of geography and urbanism. The data gather the recording of daily measurements of five environmental variables, namely three pollutants (PM10, NO2, O3) and two meteorological factors (pressure and temperature) over six years. Those data can be seen as multivariate functional data: quantitative entities evolving along time, for which there is a growing need of methods to summarize and understand them. For this purpose, a novel co-clustering model for multivariate functional data is defined. The model is based on a functional latent block model which assumes for each co-cluster a probabilistic distribution for multivariate functional principal component scores. A Stochastic EM algorithm, embedding a Gibbs sampler, is proposed for model inference, as well as a model selection criteria for choosing the number of co-clusters. The application of the proposed co-clustering algorithm on environmental data of the Région Sud allowed to divide the region composed by 357 zones in six macro-areas with common exposure to pollution. We showed that pollution profiles vary accordingly to the seasons and the patterns are conserved during the 6 years studied. These results can be used by local authorities to develop specific programs to reduce pollution at the macro-area level and to identify specific periods of the year with high pollution peaks in order to set up specific prevention programs for health. Overall, the proposed co-clustering approach is a powerful resource to analyse multivariate functional data in order to identify intrinsic data structure and summarize variables profiles over long periods of time. Figure 1 illustrates the spatial and temporal clustering results.

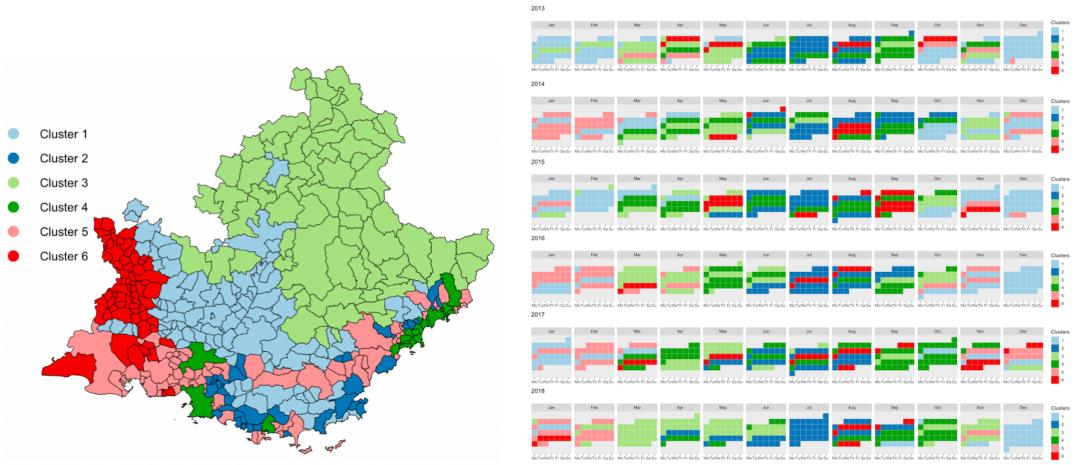


Figure 1: Spatial clustering of the area zones according to the air pollution dynamic over the studied period (left panel) and temporal segmentation of the time (right panel). Those tools may offer meaningful summaries on such massive pollution data to experts or local authorities.

7.1.4 Semi-supervised Consensus Clustering Based on Closed Patterns

Participants: Frédéric Precioso.

Keywords: Clustering; Semi-supervised learning; Semi-supervised consensus clustering; Frequent closed itemsets
Collaborations: Tianshu YANG (Université Côte d'Azur, Amadeus), Nicolas PASQUIER (Université Côte d'Azur), Luca MARCHETTI (Amadeus), Michael DEFOIN PLATEL (Amadeus), in a CIFRE PhD project with Amadeus

Semi-supervised consensus clustering, also called semi-supervised ensemble clustering, is a recently emerged technique that integrates prior knowledge into consensus clustering in order to improve the quality of the clustering result. In this article [25], we propose a novel semi-supervised consensus clustering algorithm extending the previous work on the MultiCons multiple consensus clustering approach. By using closed pattern mining technique, the proposed Semi-MultiCons algorithm manages to generate a recommended consensus solution with a relevant inferred number of clusters based on ensemble members with different and pairwise constraints. Compared with other semi-supervised and/or consensus clustering approaches, Semi-MultiCons does not require the number of generated clusters as an input parameter, and is able to alleviate the widely reported negative effect related to the integration of constraints into clustering. The experimental results demonstrate that the proposed method outperforms state of the art semi-supervised consensus clustering algorithms.

7.1.5 Dimension-Grouped Mixed Membership Models for Multivariate Categorical Data

Participants: Elena Erosheva.

Keywords: Bayesian estimation, grant peer review, inter-rater reliability, maximum likelihood estimation, measurement, mixed-effects models
Collaborations: Yuqi Gu (Columbia University), Gongjun Xu (University of Michigan), David B. Dunson (Duke University)

Mixed Membership Models (MMMs) are a popular family of latent structure models for complex multivariate data. Instead of forcing each subject to belong to a single cluster, MMMs incorporate a vector

of subject-specific weights characterizing partial membership across clusters. With this flexibility come challenges in uniquely identifying, estimating, and interpreting the parameters. In [40], we propose a new class of Dimension-Grouped MMMs (Gro-M^{3}s) for multivariate categorical data, which improve parsimony and interpretability. In Gro-M^{3}s, observed variables are partitioned into groups such that the latent membership is constant for variables within a group but can differ across groups. Traditional latent class models are obtained when all variables are in one group, while traditional MMMs are obtained when each variable is in its own group. The new model corresponds to a novel decomposition of probability tensors. Theoretically, we derive transparent identifiability conditions for both the unknown grouping structure and model parameters in general settings. Methodologically, we propose a Bayesian approach for Dirichlet Gro-M3 s to inferring the variable grouping structure and estimating model parameters. Simulation results demonstrate good computational performance and empirically confirm the identifiability results. We illustrate the new methodology through an application to a functional disability dataset.

7.1.6 Hierarchical clustering with discrete latent variable models and the ICL criterion

Participants: Charles Bouveyron.

Keywords: generative models, model-based clustering, model selection, discrete latent variable models, networks
Collaborations: Pierre Latouche (Univ. de Paris), Nicolas Jouvin (INRAE & AgroParisTech), E. Côme (Univ. Gustave Eiffel)

In [16], we introduce a two step methodology to extract a hierarchical clustering. This methodology considers the integrated classification likelihood criterion as an objective function, and applies to any discrete latent variable models (DLVM) where this quantity is tractable. The first step of the methodology involves maximizing the criterion with respect to the discrete latent variables state with uninformative priors. To that end we propose a new hybrid algorithm based on greedy local searches as well as a genetic algorithm which allows the joint inference of the number K of clusters and of the clusters themselves. The second step of the methodology is based on a bottom-up greedy procedure to extract a hierarchy of clusters from this natural partition. In a Bayesian context, this is achieved by considering the Dirichlet cluster proportion prior parameter α as a regularisation term controlling the granularity of the clustering. This second step allows the exploration of the clustering at coarser scales and the ordering of the clusters an important output for the visual representations of the clustering results. The clustering results obtained with the proposed approach, on simulated as well as real settings, are compared with existing strategies and are shown to be particularly relevant. This work is implemented in the R package `greed` and Figure 2 illustrates the main idea of the method.

7.1.7 Tensor decomposition for learning Gaussian mixtures from moments

Participants: Pierre-Alexandre Mattei.

Keywords: model-based clustering, tensor decomposition, method of moments
Collaborations: Rima Khouja, Bernard Mourrain (Inria Sophia-Antipolis, AROMATH team)

In [41] consider the problem of estimation of Gaussian mixture models. As an alternative to maximum-likelihood, our focus is on the method of moments. More specifically, we investigate symmetric tensor decomposition methods, where the tensor is built from empirical moments of the data distribution. We consider identifiable tensors, which have a unique decomposition, showing that moment tensors built from spherical Gaussian mixtures have this property. We prove that symmetric tensors with interpolation degree strictly less than half their order are identifiable and we present an algorithm, based on simple linear algebra operations, to compute their decomposition. Illustrative experimentations show the

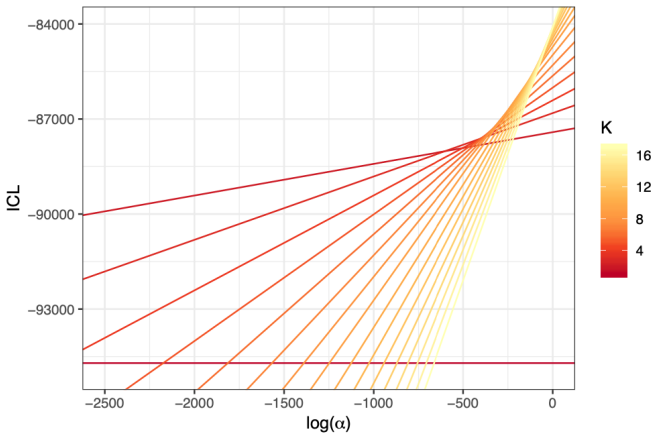


Figure 2: Lines of slope representing the ICL function according to log(α) for collections of partitions with a decreasing number of hierarchical clusters.

impact of the tensor decomposition method for recovering Gaussian mixtures, in comparison with other state-of-the-art approaches.

7.1.8 Co-clustering of time-dependent data via a shape invariant model, with application to the modeling of COVID-19 evolution across countries

Participants: Charles Bouveyron, Elena Erosheva.

Keywords: Model-based coclustering, time-dependent data, latent block model
Collaborations: Alessandro Casa (University College Dublin, Ireland), Giovanna Menardi (Universita degli Studi di Padova, Italy)

Multivariate time-dependent data, where multiple features are observed over time for a set of individuals, are increasingly widespread in many application domains. To model these data, we need to account for relations among both time instants and variables and, at the same time, for subject heterogeneity. We propose in [13] a new co-clustering methodology for grouping individuals and variables simultaneously, designed to handle both functional and longitudinal data. Our approach borrows some concepts from the curve registration framework by embedding the shape invariant model in the latent block model, estimated via a suitable modification of the SEM-Gibbs algorithm. The resulting procedure allows for several user-defined specifications of the notion of cluster that can be chosen on substantive grounds and provides parsimonious summaries of complex time-dependent data by partitioning data matrices into homogeneous blocks. Along with the explicit modelling of time evolution, these aspects allow for an easy interpretation of the clusters, from which also low-dimensional settings may benefit. The proposed method is applied to the modeling of COVID-19 evolution across countries. Figure 3 illustrates the kind of data this work considers.

7.1.9 Comparison-based centrality measures

Participants: Damien Garreau.

Collaborations: Luca Rendsburg (University of Tübingen)

Recently, learning only from ordinal information of the type "item x is closer to item y than to item z" has received increasing attention in the machine learning community. Such triplet comparisons are

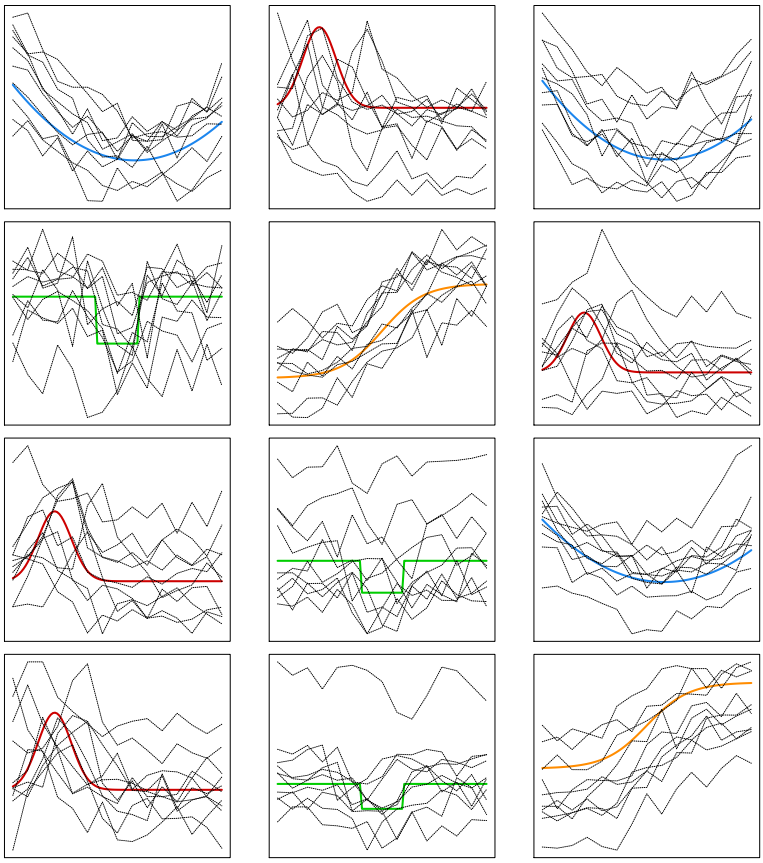


Figure 3: Co-clustering of time-dependent data via a shape invariant model.

particularly well suited for learning from crowdsourced human intelligence tasks, in which workers make statements about the relative distances in a triplet of items. In [23], we systematically investigate comparison-based centrality measures on triplets and theoretically analyze their underlying Euclidean notion of centrality. Two such measures already appear in the literature under opposing approaches, and we propose a third measure, which is a natural compromise between these two. We further discuss their relation to statistical depth functions, which comprise desirable properties for centrality measures, and conclude with experiments on real and synthetic datasets for medoid estimation and outlier detection. Figure 4 illustrates this work.

7.1.10 Dynamic Co-Clustering for PharmaCovigilance

Participants: Charles Bouveyron, Marco Corneli, Giulia Marchello.

Keywords: generative models, dynamic co-clustering, count data, pharmacovigilance
Collaborations: Audrey Fresse (Centre de Pharmacovigilance, CHU de Nice)

We consider in [43] the problem of co-clustering count matrices with a high level of missing values that may evolve in time. We introduce a generative model, named dynamic latent block model (dLBM), which extends the classical binary latent block model (LBM) to the dynamic case. The time dependent counting data are modeled via non-homogeneous Poisson processes (HHPPs). The continuous time is handled by a partition of the whole considered time period, with the interaction counts being aggregated on the time intervals of such partition. In this way, a sequence of static matrices that allows us to identify

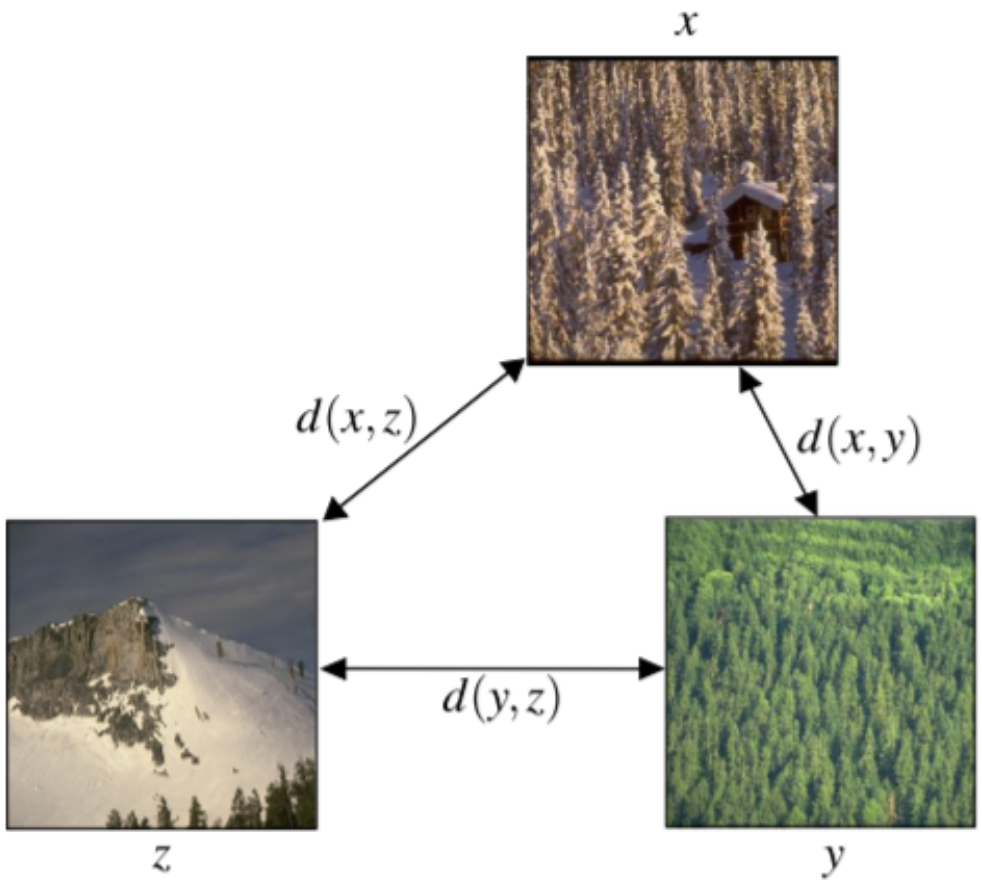


Figure 4: The triplet setting for comparison-based centrality measures.

meaningful time clusters is obtained. The model inference is done using a SEM-Gibbs algorithm and the ICL criterion is used for model selection. Numerical experiments on simulated data highlight the main features of the proposed approach and show the interest of dLBM with respect to related works. An application to adverse drug reaction (ADR) dataset, obtained thanks to the collaboration with the Regional Center of Pharmacovigilance (RCPV) of Nice (France), is also proposed. One of the missions of RCPVs is safety signal detection. However, the current expert detection of safety signals, despite being unavoidable, has the disadvantage of being incomplete due to the workload it represents. For this reason, developing automatized method of safety signal detection is currently a major issue in pharmacovigilance. The application of dLBM on this dataset allowed us to extract meaningful patterns for medical authorities. In particular, dLBM identifies 7 drug clusters, 10 ADRs clusters and 6 time clusters. The clusters identified by the algorithm are coherent with previous knowledge and adequately represent the variety of drugs present in the dataset. Moreover, an in depth analysis of the clusters found by the model revealed that dLBM correctly detected the three drugs that gave rise to the health scandals that took place between 2010 and 2020, demonstrating its potential as a routine tool in pharmacovigilance. Figure 5 illustrates this work.

7.1.11 Bayesian discriminative Gaussian clustering

Participants: Charles Bouveyron.

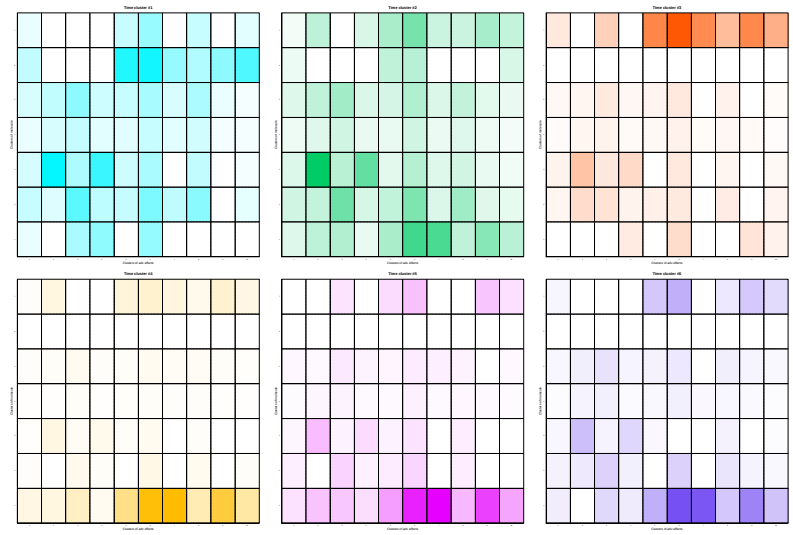


Figure 5: Evolution of the relation between the drug clusters and the ADR clusters over time. Each color corresponds to a different cluster of adverse drug reaction.

Keywords: generative models, model-based clustering, Bayesian modeling, high-dimensional data,
Collaborations: Pierre Latouche (Univ. de Paris), Nicolas Jouvin (INRAE & AgroParisTech)

High-dimensional data clustering has become and remains a challenging task for modern statistics and machine learning, with a wide range of applications. We considered in [20] the powerful discriminative latent mixture model, and we extended it to the Bayesian framework. Modeling data as a mixture of Gaussians in a low-dimensional discriminative subspace, a Gaussian prior distribution is introduced over the latent group means and a family of twelve submodels are derived considering different covariance structures. Model inference is done with a variational EM algorithm, while the discriminative subspace is estimated via a Fisher-step maximizing an unsupervised Fisher criterion. An empirical Bayes procedure is proposed for the estimation of the prior hyper-parameters, and an integrated classification likelihood criterion is derived for selecting both the number of clusters and the submodel. The performances of the resulting Bayesian Fisher-EM algorithm are investigated in two thorough simulated scenarios, regarding both dimensionality as well as noise and assessing its superiority with respect to state-of-the-art Gaussian subspace clustering models. In addition to standard real data benchmarks, an application to single image denoising is proposed, displaying relevant results. This work comes with a reference implementation for the R software in the `FisherEM` package accompanying the paper. Figure 6 illustrates this work.

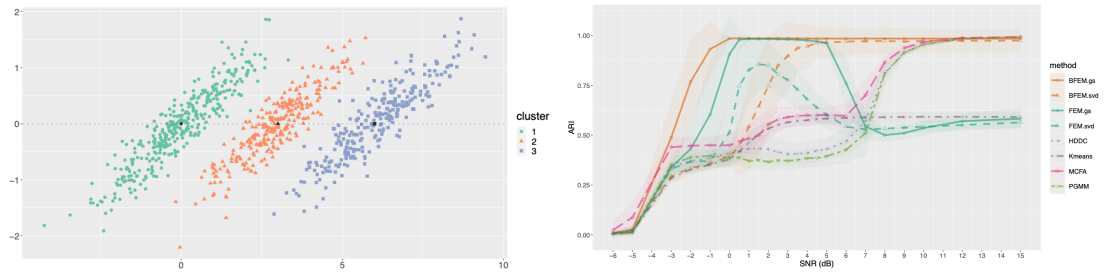


Figure 6: Comparison of the clustering performance of BFEM compared to competitive methods for (simulated) high-dimensional data. Numerical experiments showed that BFEM outperforms competitors in most situations, and in particular in difficult situations where the noise-to-signal ratio is very high.

7.1.12 Unsupervised classification of SDSS galaxy spectra

Participants: Charles Bouveyron.

Keywords: generative models, model-based clustering, high-dimensional data, astrophysique
Collaborations: D. Fraix-Burnet (Univ. Grenoble Alpes & CNRS), J. Moultaka (CNRS)

In this applied work [19], we use the Fisher-EM algorithm for clustering for the unsupervised classification of 702, 248 spectra of galaxies and quasars with resdshifts smaller than 0.25 that were retrieved from the Sloan Digital Sky Survey (SDSS) database, release 7. The spectra were first corrected for the redshift, then wavelet-filtered to reduce the noise, and finally binned to obtain about 1437 wavelengths per spectrum. Fisher-EM, an unsupervised clustering discriminative latent mixture model algorithm, was applied on these corrected spectra, considering the full set as well as several subsets of 100,000 and 300,000 spectra. The optimum number of classes given by a penalized likelihood criterion is 86 classes, the 37 most populated ones gathering 99% of the sample. These classes are established from a subset of 302144 spectra. Using several cross-validation techniques we find that this classification is in agreement with the results obtained on the other subsets with an average misclassification error of about 15%. The large number of very small classes tends to increase this error rate. This is the first time that an automatic, objective and robust unsupervised classification is established on such a large amount of spectra of galaxies. The mean spectra of the classes can be used as templates for a large majority of galaxies in our Universe. Figure 7 illustrates the obtained results.

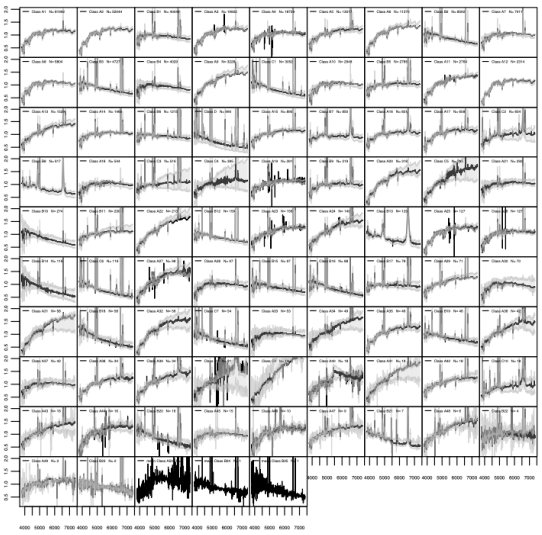


Figure 7: Clustering results with Fisher-EM of the 702, 248 spectra of galaxies and quasars, into 86 clusters.

7.2 Understanding (deep) learning models

7.2.1 Multi-dimensional text analysis through deep networks

Participants: Frédéric Precioso, Laurent Vanni.

Keywords: Text classification, Convolutional Neural Networks, Recurrent Neural Networks, Deep linguistic patterns

Collaborations: Damon Mayaffre (CNRS)

Historically, text mining has gone from processing raw texts to processing labelled texts. If the graphic form of a word is often sufficient, there are many ambiguous cases of homonyms where the use of a simple dictionary of forms is not sufficient. Lemmatization is an efficient way to remove these ambiguities and obtain a description of words on several levels of representation. Moreover, lemmatizers usually allow to associate a morphosyntactic label (part of speech, verbal tense, gender, number) to the word. It is this linguistic information that becomes valuable for our descriptive approach through deep learning, as it allows us to observe complex lexico-grammatical structures, that potentially associate several levels of text representation in the same structure. The convolutional model used until now must therefore be adapted to integrate this additional information in order to obtain an even finer description of the textual salience of a corpus.

In this book chapter [36], we recap the potential of deep networks for mutlidimensional text analysis, leading us to define the new concept of deep linguistic patterns.

7.2.2 From text saliency to linguistic objects: learning linguistic interpretable markers with a multi-channels convolutional architecture

Participants: Marco Corneli, Frédéric Precioso, Laurent Vanni.

Keywords: Deep Learning, Interpretation, Linguistic, Text analysis
Collaborations: Damon Mayaffre (CNRS)

A lot of effort is currently made to provide methods to analyze and understand deep neural network impressive performances for tasks such as image or text classification. These methods are mainly based on visualizing the important input features taken into account by the network to build a decision. However these techniques, let us cite LIME, SHAP, Grad-CAM, or TDS, require extra effort to interpret the visualization with respect to expert knowledge. In this work [45], we propose a novel approach to inspect the hidden layers of a fitted CNN in order to extract interpretable linguistic objects from texts exploiting classification process. In particular, we detail a weighted extension of the Text Deconvolution Saliency (wTDS) measure which can be used to highlight the relevant features used by the CNN to perform the classification task. We empirically demonstrate the efficiency of our approach on corpora from two different languages: English and French. On all datasets, wTDS automatically encodes complex linguistic objects based on co-occurrences and possibly on grammatical and syntax analysis.

7.2.3 Learning and Reasoning for Cultural Metadata Quality

Participants: Frédéric Precioso.

Keywords: Deep Learning, Image Recognition, Semantic Web, Knowledge Graph
Collaborations: Anna Bobasheva, Fabien Gandon (Inria)

This work [12] combines semantic reasoning and machine learning to create tools that allow curators of the visual art collections to identify and correct the annotations of the artwork as well as to improve the relevance of the content-based search results in these collections. The research is based on the Joconde database maintained by French Ministry of Culture that contains illustrated artwork records from main French public and private museums representing archeological objects, decorative arts, fine arts, historical and scientific documents, etc. The Joconde database includes semantic metadata that describes properties of the artworks and their content. The developed methods create a data pipeline that processes metadata, trains a Convolutional Neural Network image classification model, makes prediction for the entire collection and expands the metadata to be the base for the SPARQL search queries. We developed a set of such queries to identify noise and silence in the human annotations and to search image content with results ranked according to the relevance of the objects quantified by the prediction score provided by the deep learning model. We also developed methods to discover new contextual

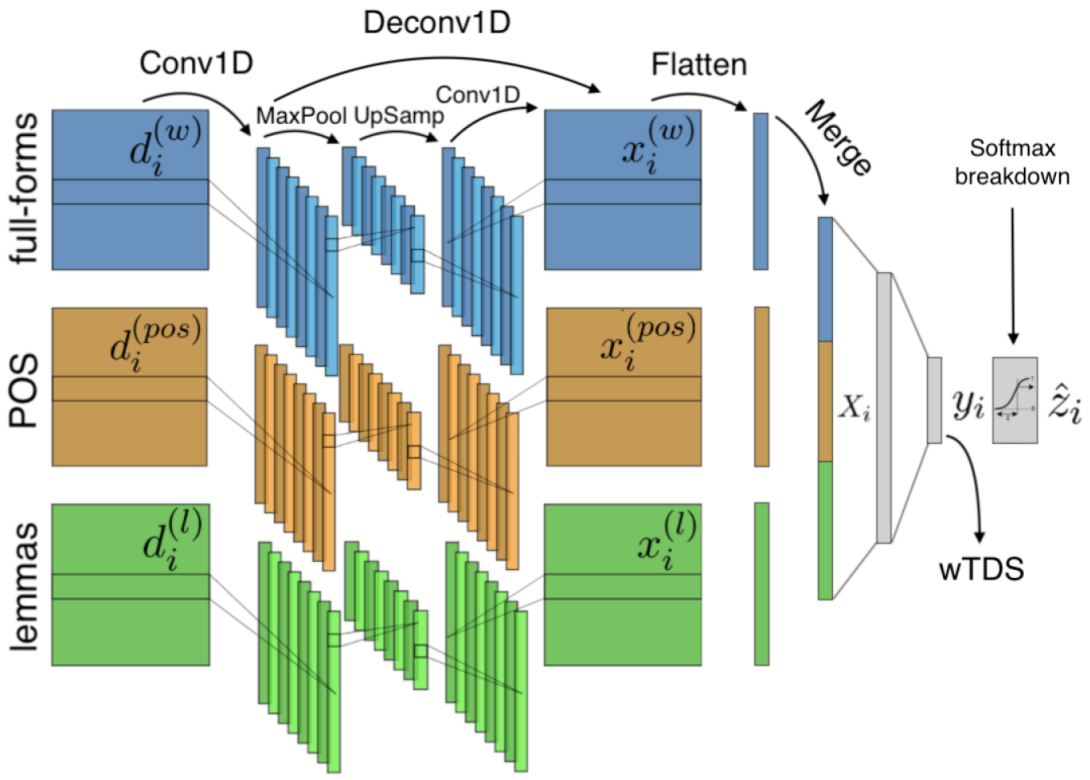


Figure 8: Three channels convolution/deconvolution for three representation of the input 1) full-forms (words), 2) part-of-speech (POS), 3) lemma.

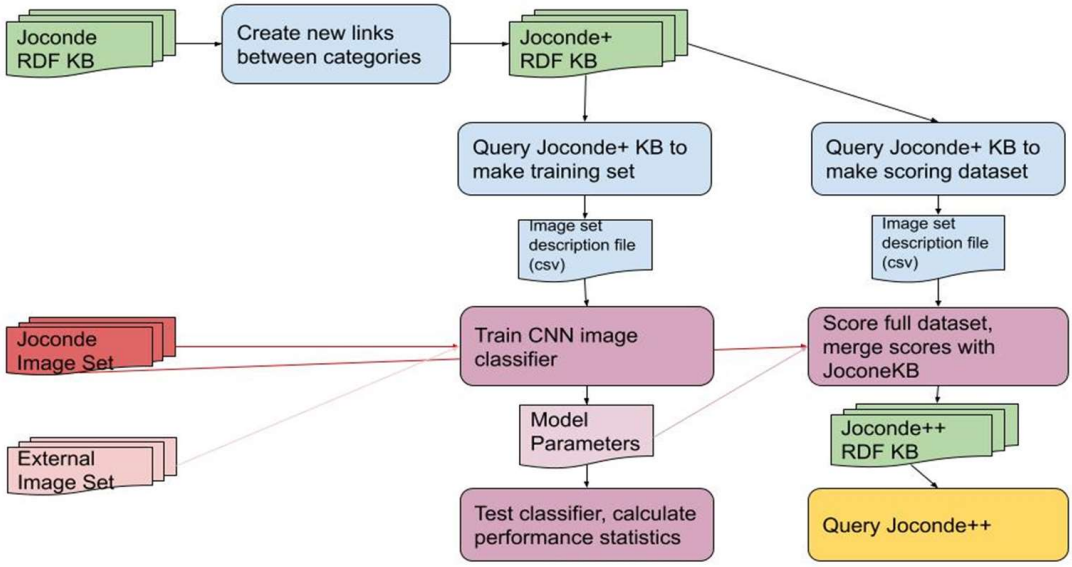


Figure 9: Data processing pipeline combining reasoning and learning.

relationships between the concepts in the metadata by analyzing the contrast between the concepts similarities in the Joconde's semantic model and other vocabularies and we tried to improve the model prediction scores based on the semantic relations. Our results show that cross-fertilization between symbolic AI and machine learning can indeed provide the tools to address the challenges of the museum curators work describing the artwork pieces and searching for the relevant images.

7.2.4 A New Method from a Re-examination of Deep Architectures for Head Motion Prediction in 360° Videos

Participants: Frédéric Precioso.

Keywords: Modeling and prediction, Virtual reality, Neural nets, Machine Learning, Kinematics and dynamics
Collaborations: Miguel Fabian Romero Rondon, Lucile Sassatelli, Ramon Aparicio Pardo (I3S)

In this work [24], we consider predicting the user's head motion in 360° videos, with 2 modalities only: the past user's positions and the video content (not knowing other users' traces). We make two main contributions. First, we re-examine existing deep-learning approaches for this problem and identify hidden flaws from a thorough root-cause analysis. Second, from the results of this analysis, we design a new proposal establishing state-of-the-art performance. First, re-assessing the existing methods that use both modalities, we obtain the surprising result that they all perform worse than baselines using the user's trajectory only. A root-cause analysis of the metrics, datasets and neural architectures shows in particular that (i) the content can inform the prediction for horizons longer than 2 to 3 sec. (existing methods consider shorter horizons), and that (ii) to compete with the baselines, it is necessary to have a recurrent unit dedicated to process the positions, but this is not sufficient. Second, from a re-examination of the problem supported with the concept of Structural-RNN, we design a new deep neural architecture, named TRACK. TRACK achieves state-of-the-art performance on all considered datasets and prediction horizons, outperforming competitors by up to 20% on focus-type videos and horizons 2-5

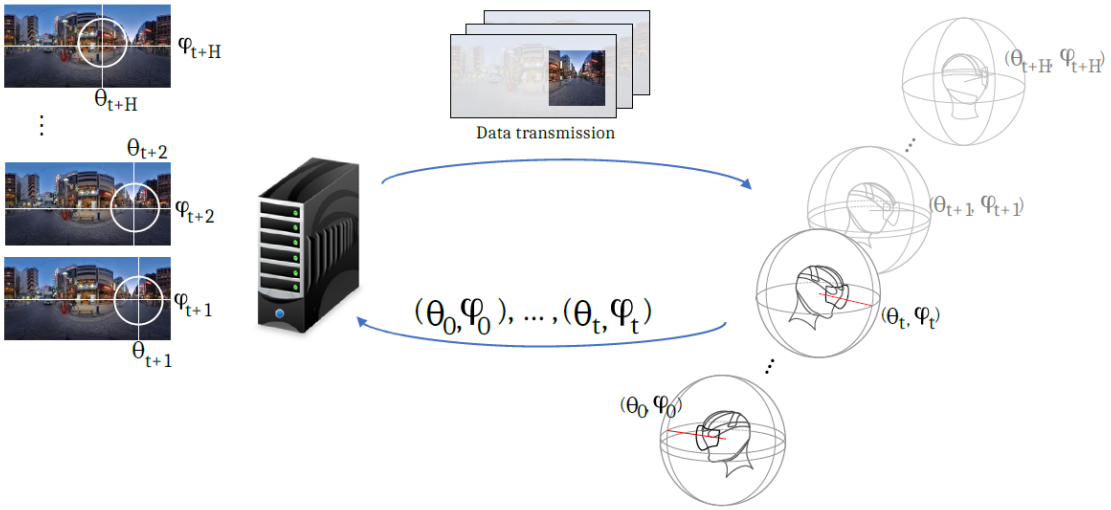


Figure 10: 360° video streaming principle. The user requests the next video segment at time t, if the future orientations of the user $(\theta_{t+1},\phi_{t+1})$,..., $(\theta_{t+H},\phi_{t+H})$ were known, the bandwidth consumption could be reduced by sending in higher quality only the areas corresponding to the future FoV.

seconds. The entire framework (codes and datasets) is online and received an ACM reproducibility badge https://gitlab.com/miguelfromeror/head-motion-prediction.

7.2.5 An analysis of LIME for text data

Participants: Damien Garreau.

Keywords: machine learning, interpretability, statistical learning theory, natural language processing
Collaborations: Dina Mardaoui (Polytech Nice Sophia, UCA)

Text data are increasingly handled in an automated fashion by machine learning algorithms. But the models handling these data are not always well-understood due to their complexity and are more and more often referred to as "black-boxes." Interpretability methods aim to explain how these models operate. Among them, LIME has become one of the most popular in recent years. However, it comes without theoretical guarantees: even for simple models, we are not sure that LIME behaves accurately. In [28], we provide a first theoretical analysis of LIME for text data. As a consequence of our theoretical findings, we show that LIME indeed provides meaningful explanations for simple models, namely decision trees and linear models. Figure 11 illustrates the way LIME explains a prediction.

7.2.6 What does LIME really see in images

Participants: Damien Garreau.

Keywords: machine learning, interpretability, statistical learning theory, computer vision
Collaborations: Dina Mardaoui (Polytech Nice Sophia, UCA)

The performance of modern algorithms on certain computer vision tasks such as object recognition is now close to that of humans. This success was achieved at the price of complicated architectures depending on millions of parameters and it has become quite challenging to understand how particular predictions are made. Interpretability methods propose to give us this understanding. In [26], we

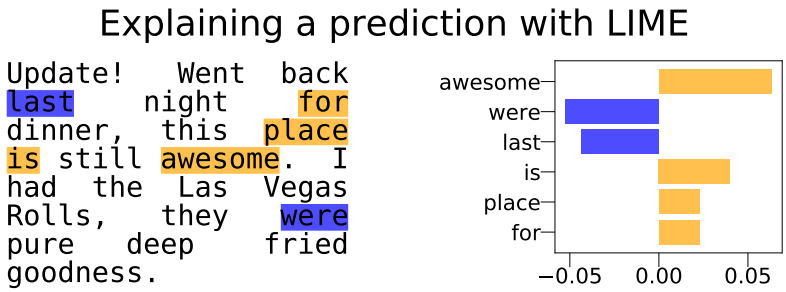


Figure 11: Explaining a prediction with LIME

extended the analysis of [28] to the image case. On the theoretical side, we show that when the number of generated examples is large, LIME explanations are concentrated around a limit explanation for which we give an explicit expression. We further this study for elementary shape detectors and linear models. As a consequence of this analysis, we uncover a connection between LIME and integrated gradients, another explanation method. More precisely, the LIME explanations are similar to the sum of integrated gradients over the superpixels used in the preprocessing step of LIME. Figure 12 illustrates the work.

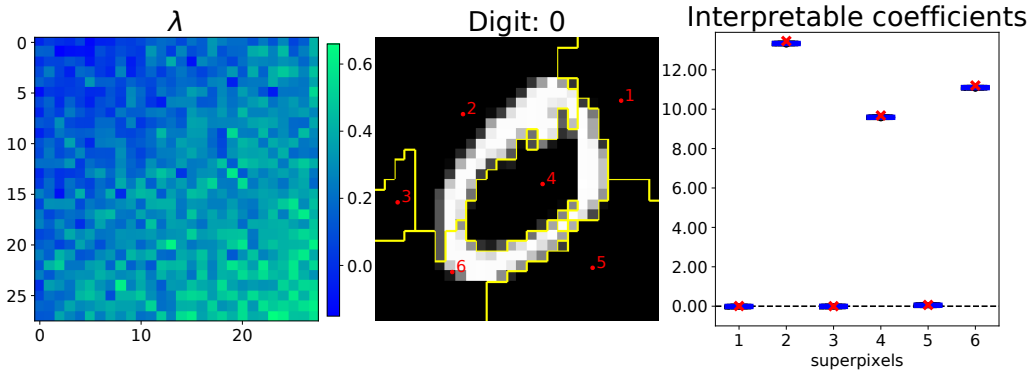


Figure 12: Theory vs practice on an arbitrary linear model, MNIST image, LIME default settings.E

7.2.7 SMACE: A New Method for the Interpretability of Composite Decision Systems

Participants: Damien Garreau, Gianluigi Lopardo, Greger Ottosson, Frédéric Precioso.

Keywords: Interpretability, Explainable AI, Composite AI, Machine Learning, Decision-Making

Interpretability is a pressing issue for decision systems. Many post hoc methods have been proposed to explain the predictions of any machine learning model. However, business processes and decision systems are rarely centered around a single, standalone model. These systems combine multiple models that produce key predictions, and then apply decision rules to generate the final decision. To explain such decision, we presented in [42] SMACE, Semi-Model-Agnostic Contextual Explainer, a novel interpretability method that combines a geometric approach for decision rules with existing post hoc solutions for machine learning models to generate an intuitive feature ranking tailored to the end user. We show that established model-agnostic approaches produce poor results in this framework. Figure 13 illustrates this work.

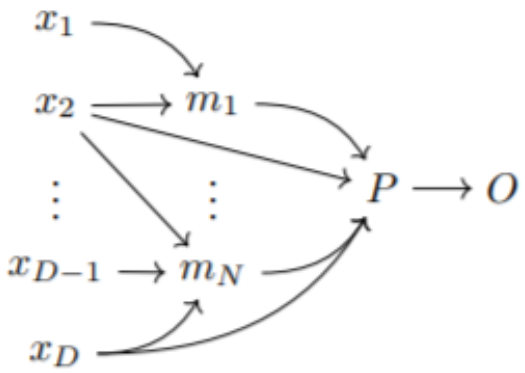


Figure 13: Structure of a composite decision-making system with D input features $x_1,\ldots,x_D$, and N models $m_1,\ldots,m_N$. A decision policy P (*i.e.*, a set decision of rules) is finally applied to produce an outcome O. Note that in general both the models and the rules take a subset of input features as input, tough not necessarily the same.

7.3 Adaptive and robust learning

7.3.1 Unobserved classes and extra variables detection in high-dimensional discriminant analysis

Participants: Charles Bouveyron, Pierre-Alexandre Mattei.

Keywords: Adaptive supervised classification; conditional estimation; model-based discriminant analysis; unobserved classes; variable selection.
Collaborations: Micheal Fop and Brendan Murphy (University College Dublin, Ireland)

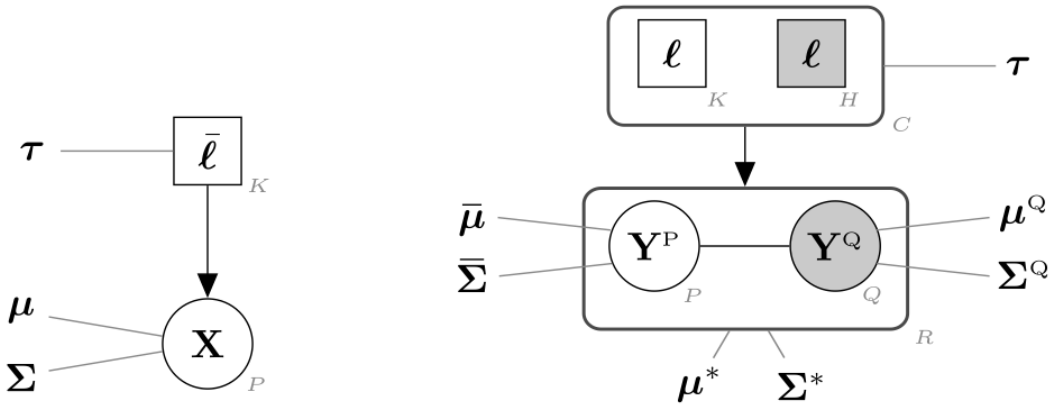
In supervised classification problems, the test set may contain data points belonging to classes not observed in the learning phase. Moreover, the same units in the test data may be measured on a set of additional variables recorded at a subsequent stage with respect to when the learning sample was collected. In this situation, the classifier built in the learning phase needs to adapt to handle potential unknown classes and the extra dimensions. We introduce in [18] a model-based discriminant approach, Dimension-Adaptive Mixture Discriminant Analysis (D-AMDA), which can detect unobserved classes and adapt to the increasing dimensionality. Model estimation is carried out via a full inductive approach based on an EM algorithm. The method is then embedded in a more general framework for adaptive variable selection and classification suitable for data of large dimensions. A simulation study and an artificial experiment related to classification of adulterated honey samples are used to validate the ability of the proposed framework to deal with complex situations. Figure 14 illustrates the general framework of the proposed approach.

7.3.2 Knowledge-Driven Active Learning

Participants: Gabriele Ciravegna, Marco Gori, Frédéric Precioso.

Keywords: Active Learning, Knowledge Representation, Deep Learning

Deep Learning (DL) methods have achieved impressive results over the last few years in fields ranging from computer vision to machine translation [56]. Most of the research, however, focused on improving model performances, while little attention has been paid to overcome the intrinsic limits of DL algorithms. In particular, in this work [38] we will focus on the amount of data problem. Indeed, deep neural networks



(a) *Learning phase.* (b) *Discovery phase.*

Figure 14: General framework of the inductive approach for Dimension-Adaptive Mixture Discriminant Analysis (DAMDA).

need large amounts of labelled data to be properly trained. With the advent of Big Data, sample collection does not represent an issue any more. Nonetheless, the number of supervised data in some contexts is limited, and manual labelling can be expensive and time-consuming. Therefore, a common situation is the unlabelled pool scenario, where many data are available, but only some are annotated. Historically, two strategies have been devised to tackle this situation: semi-supervised learning which focus on improving feature representations by processing the unlabelled data with unsupervised techniques; active learning in which the training algorithm indicates which data should be annotated to improve the most its performances. The main assumption behind active learning strategies is that there exists a subset of samples that allows to train a model with a similar accuracy as when fed with all training data. Iteratively, the model indicates the optimal samples to be annotated from the unlabelled pool. This is generally done by ranking the unlabelled samples w.r.t. a given measure and by selecting the samples associated to the highest scores. In this paper, we propose an active learning strategy that compares the predictions over the unsupervised data with the available domain knowledge and exploits the inconsistencies as an index for selecting the data to be annotated. Domain knowledge can be generally expressed as First-Order Logic (FOL) clauses and translated into real-valued logic constraints by means of T-Norms. This formulation has been employed in the semi-supervised learning scenario to improve classifier performance by enforcing the constraints on the unsupervised data. More recently, constraints violation has been effectively used also as a metric to detect adversarial attacks. To the best of our knowledge, however, domain-knowledge (in the form of logic constraints) violation has never been used as an index in the selection process of an active learning strategy. We show that the proposed strategy outperforms the standard uncertain sample selection method, particularly in those contexts where domain-knowledge is rich. We empirically demonstrate that this is mainly due to the fact that the proposed strategy allows discovering data distributions lying far from training data, unlike uncertainty-based approaches. Neural networks, indeed, are known to be over-confident of their prediction, and they are generally unable to recognize samples lying far from the training data distribution. This issue, beyond exposing them to adversarial attacks, prevents uncertainty-based strategies from detecting these samples as points that would require an annotation. On the contrary, even though a neural network may be confident of its predictions, the interaction between the predicted classes may still offer a way to spot out-of-the-distribution samples. Finally, the Knowledge-driven Active Learning (KAL) strategy can be also employed in the object-detection context where standard uncertainty-based ones are difficult to apply.

7.3.3 Kernel-Matrix Determinant Estimates from stopped Cholesky Decomposition

Participants: Damien Garreau.

Keywords: Gaussian processes, Cholesky decomposition, kernel matrix
Collaborations: Simon Bartels (University of Copenhaguen), Wouter Boosma (University of Copenhaguen), Jes Frellsen (Technical University of Denmark)

Algorithms involving Gaussian processes or determinantal point processes typically require computing the determinant of a kernel matrix. Frequently, the latter is computed from the Cholesky decomposition, an algorithm of cubic complexity in the size of the matrix. We show that, under mild assumptions, it is possible to estimate the determinant from only a sub-matrix, with probabilistic guarantee on the relative error. In [37], we present an augmentation of the Cholesky decomposition that stops under certain conditions before processing the whole matrix. Experiments demonstrate that this can save a considerable amount of time while having an overhead of less than 5% when not stopping early. More generally, we present a probabilistic stopping strategy for the approximation of a sum of known length where addends are revealed sequentially. We do not assume independence between addends, only that they are bounded from below and decrease in conditional expectation.

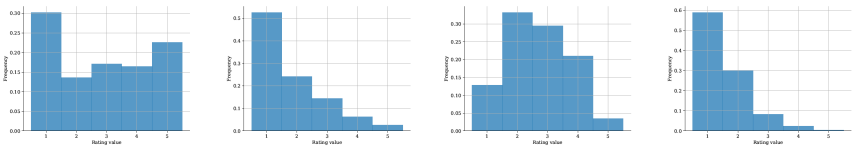
7.4 Learning with heterogeneous and corrupted data

7.4.1 Deep generative modelling with missing not at random data

Participants: Pierre-Alexandre Mattei.

Keywords: missing data, neural networks, deep learning, generative models,
Collaborations: Jes Frellsen, Niel Bruun Ipsen (Technical University of Denmark)

When a missing process depends on the missing values themselves, it needs to be explicitly modelled and taken into account while doing likelihood-based inference. We present an approach for building and fitting deep latent variable models (DLVMs) in cases where the missing process is dependent on the missing data. Specifically, a deep neural network enables us to flexibly model the conditional distribution of the missingness pattern given the data. This allows for incorporating prior information about the type of missingness (e.g. self-censoring) into the model. Our inference technique, based on importance-weighted variational inference, involves maximising a lower bound of the joint likelihood. Stochastic gradients of the bound are obtained by using the reparameterisation trick both in latent space and data space. Our method is called *not-missing-at-random importance-weighted autoencoder (not-MIWAE)* [27]. We show on various kinds of data sets and missingness patterns that explicitly modelling the missing process can be invaluable. We apply our method to censoring in datasets from the UCI database, clipping in images and the issue of selection bias in recommender systems (see Fig. 15).



(a) MNAR train samples (b) MCAR test samples (c) MIWAE impute (d) not-MIWAE impute

Figure 15: Using not-MIWAE to tackle selection bias in a recommender system. Yahoo! Histograms over rating values from (a) the MNAR training set and (b) the MCAR test set. (c) and (d) show histograms over imputations of missing values in the test set, when encoding the corresponding training set. The not-MIWAE imputations (d) are much more faithful to the shape of the test set (b) than the MIWAE imputations (c).

7.4.2 Active Speaker Detection as a Multi-Objective Optimization with Uncertainty-based Multimodal Fusion

Participants: Charles Bouveyron, Baptiste Pouthier, Frédéric Precioso.

Keywords: Active speaker detection, Audiovisual, Multimodal fusion, Multi-objective.
Collaborations: Laurent Pilati, Leela K. Gudupudi (NXP)

It is now well established from a variety of studies that there is a significant benefit from combining video and audio data in detecting active speakers. However, either of the modalities can potentially mislead audiovisual fusion by inducing unreliable or deceptive information. This work [30] outlines active speaker detection as a multi-objective learning problem to leverage best of each modalities using a novel self-attention, uncertainty-based multimodal fusion scheme. Results obtained show that the proposed multi-objective learning architecture outperforms traditional approaches in improving both mAP and AUC scores. We further demonstrate that our fusion strategy surpasses, in active speaker detection, other modality fusion methods reported in various disciplines. We finally show that the proposed method significantly improves the state-of-the-art on the AVA-ActiveSpeaker dataset.

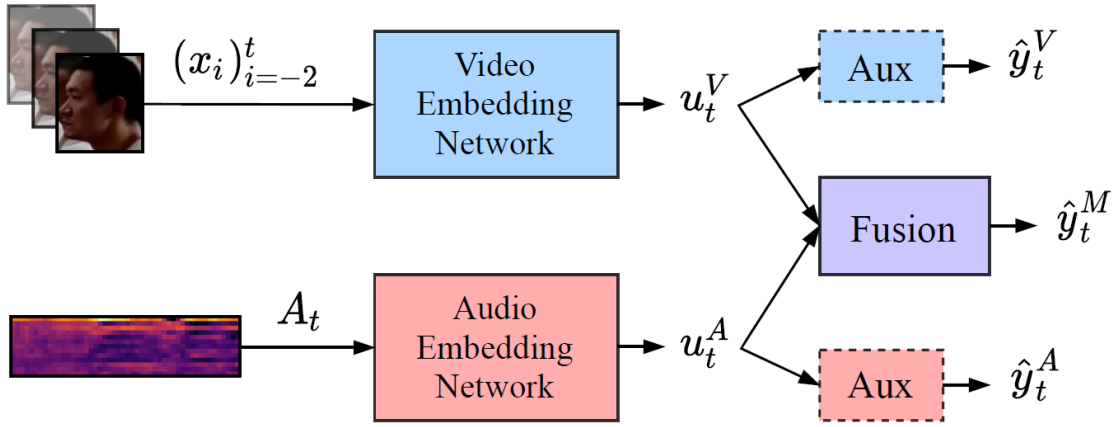


Figure 16: End-to-end multi-objective audiovisual network. "Aux" block represents an auxiliary classifier.

7.4.3 DeepLTRS: A Deep Latent Recommender System based on User Ratings and Reviews

Participants: Charles Bouveyron, Marco Corneli, Dingge Liang.

Keywords: Recommender System, Topic modelling, latent representation learning.
Collaborations: Pierre Latouche (Univ. de Paris)

We introduced in [22] a deep latent recommender system named deepLTRS in order to provide users with high quality recommendations based on observed user ratings *and* texts of product reviews. The underlying motivation is that, when a user scores only a few products, the texts used in the reviews represent a significant source of information, thereby enhancing the predictive ability of the model. Our approach adopts a variational auto-encoder (VAE) architecture as a deep generative latent model for an ordinal matrix encoding ratings and a document-term matrix encoding the reviews. Taking into account both matrices as model inputs, deepLTRS uses a neural network to capture the relationship between latent factors and latent topics. Moreover, a user-majoring encoder and a product-majoring encoder are constructed to jointly capture user and product preferences. Due to the specificity of the model structure, an original row-column alternated mini-batch optimization algorithm is proposed to

deal with user-product dependencies and computational burden. Numerical experiments on simulated and real-world data sets demonstrate that deepLTRS outperforms the state-of-the-art, in particular in context of extreme data sparsity. Figure 17 illustrates the performances achieved by deepLTRS compared to state-of-the-art methods.

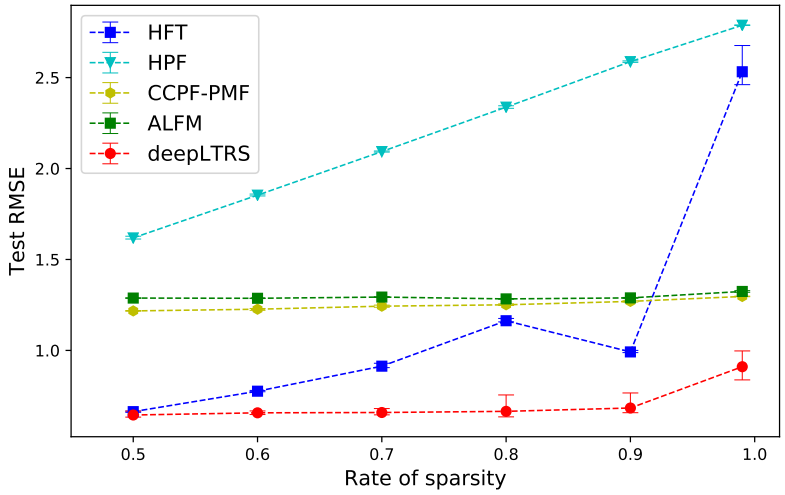


Figure 17: Test RMSE of models with different sparsity level on simulated data.

7.4.4 Hierarchical Multimodal Attention for Deep Video Summarization

Participants: Frédéric Precioso, Melissa Sanabria.

Keywords: Event stream data, Soccer match data, Video Summarization, Multimodal data, Sports Analytics
Collaborations: Thomas Menguy (Wildmoka Company), this work has been co-funded by Région Sud Provence Alpes Côte d'Azur (PACA), Université Côte d'Azur (UCA) and Wildmoka Company.

This paper explores the problem of summarizing professional soccer matches as automatically as possible using both the event-stream data collected from the field and the content broadcasted on TV. We have designed an architecture, introducing first (1) a Multiple Instance Learning method that takes into account the sequential dependency among events and then (2) a hierarchical multimodal attention layer that grasps the importance of each event in an action [31]. We evaluate our approach on matches from two professional European soccer leagues, showing its capability to identify the best actions for automatic summarization by comparing with real summaries made by human operators. Figure 18 illustrates the general schema of the approach.

7.4.5 A coherent framework for learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data

Participants: Juliette Chevallier.

Keywords: Bayesian inference, differential geometry, manifold-valued data
Collaborations: V. Debavelaere (CMAP), Stéphanie Allassonnière (Univ. de Paris)

We provide in [15] a coherent framework for studying longitudinal manifold-valued data. We introduce a Bayesian mixed-effects model which allows estimating both a group-representative piecewise-geodesic

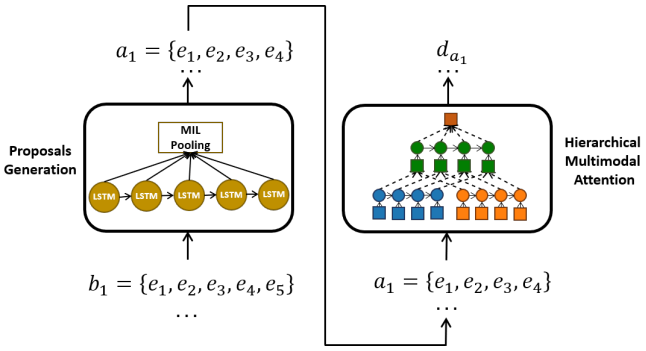


Figure 18: General schema of our approach. The left part of the figure represents the first block of our approach: Proposals Generation with a LSTM MIL Pooling. It gets as input the bags of events and outputs action Proposals. The right part of the figure is the second block of our approach: Hierarchical Multimodal Attention. It gets as input the action proposals (events data and audio data) and predict the likelihood for the given action to be in the summary.

trajectory in the Riemannian space of shape and inter-individual variability. We prove the existence of the maximum a posteriori estimate and its asymptotic consistency under reasonable assumptions. Due to the non-linearity of the proposed model, we use a stochastic version of the Expectation-Maximization algorithm to estimate the model parameters. Our simulations show that our model is not noise-sensitive and succeeds in explaining various paths of progression.

7.4.6 Online Graph Dictionary Learning

Participants: Marco Corneli, Cédric Vincent-Cuaz.

Keywords: Optimal Transport, unsupervised graph representation learning, dictionary learning.
Collaborations: Rémi Flamary, Titouan Vayer, Nicolas Courty.

Dictionary learning is a key tool for representation learning, that explains the data as linear combination of few basic elements. Yet, this analysis is not amenable in the context of graph learning, as graphs usually belong to different metric spaces. In [32], we fill this gap by proposing a new online Graph Dictionary Learning approach, which uses the Gromov-Wasserstein divergence for the data fitting term. In our work, graphs are encoded through their nodes' pairwise relations and modeled as convex combination of graph atoms, i.e. dictionary elements, estimated thanks to an online stochastic algorithm, which operates on a dataset of unregistered graphs with potentially different number of nodes. Our approach naturally extends to labeled graphs, and is completed by a novel upper bound that can be used as a fast approximation of Gromov-Wasserstein in the embedding space. We provide numerical evidences showing the interest of our approach for unsupervised embedding of graph datasets and for online graph subspace estimation and tracking. Figure 19 illustrates the approach.

7.4.7 Semi-relaxed Gromov-Wasserstein divergence with applications on graphs

Participants: Marco Corneli, Cédric Vincent-Cuaz.

Keywords: Optimal Transport, graph partitioning, dictionary learning, unsupervised graph representation learning.
Collaborations: Rémi Flamary, Titouan Vayer, Nicolas Courty.

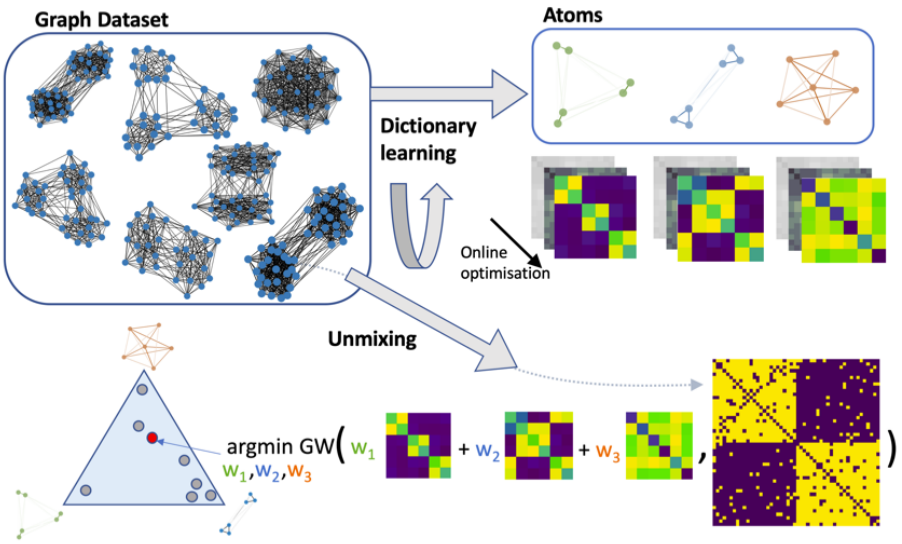


Figure 19: From a dataset of graphs with different number of nodes, our method builds a dictionary of graph atoms with an online procedure.

Comparing structured objects such as graphs is a fundamental operation involved in many learning tasks. To this end, the Gromov-Wasserstein (GW) distance, based on Optimal Transport (OT), has proven to be successful in handling the specific nature of the associated objects. More specifically, through the nodes connectivity relations, GW operates on graphs, seen as probability measures over specific spaces. At the core of OT is the idea of conservation of mass, which imposes a coupling between all the nodes from the two considered graphs. We argue in [46] that this property can be detrimental for tasks such as graph dictionary or partition learning, and we relax it by proposing a new semi-relaxed Gromov-Wasserstein divergence. Aside from immediate computational benefits, we discuss its properties, and show that it can lead to an efficient graph dictionary learning algorithm. We empirically demonstrate its relevance for complex tasks on graphs such as partitioning, clustering and completion. Figure 20 illustrates the embedding of two different graphs.

7.4.8 When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review

Participants: Elena Erosheva.

Keywords: Bayesian estimation, grant peer review, inter-rater reliability, maximum likelihood estimation, measurement, mixed-effects models
Collaborations: Patrícia Martinková (Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic), Carole J. Lee (University of Washington, Seattle)

Considerable attention has focused on studying reviewer agreement via inter-rater reliability (IRR) as a way to assess the quality of the peer review process. In [17], inspired by a recent study that reported an IRR of zero in the mock peer review of top-quality grant proposals, we use real data from a complete range of submissions to the National Institutes of Health and to the American Institute of Biological Sciences to bring awareness to two important issues with using IRR for assessing peer review quality. First, we demonstrate that estimating local IRR from subsets of restricted-quality proposals will likely result in zero estimates under many scenarios. In both data sets, we find that zero local IRR estimates are

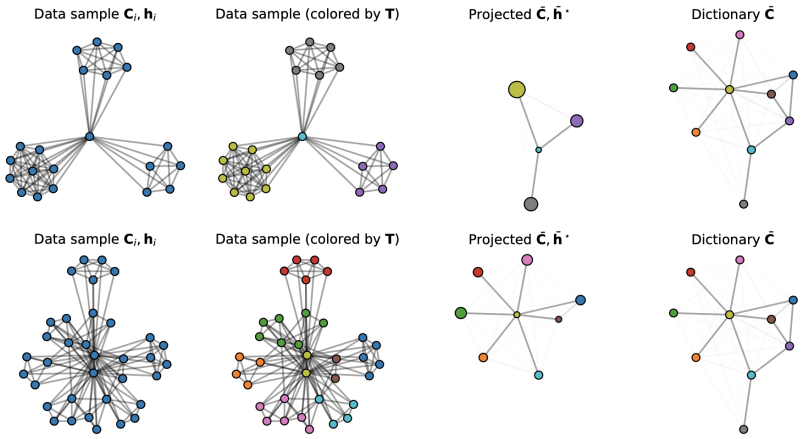


Figure 20: Illustration of the embedding of two graphs from a real dataset of social networks, on the estimated dictionary. Each row corresponds to one observed graph and we show its graph (left), its graph with nodes colored corresponding to the optimal transport plan (center left), the projected graph on the srGW dictionary with optimally reweighed nodes importance and the full dictionary with uniform weights (right).

more likely when subsets of top-quality proposals rather than bottom-quality proposals are considered. However, zero estimates from range-restricted data should not be interpreted as indicating arbitrariness in peer review. On the contrary, despite different scoring scales used by the two agencies, when complete ranges of proposals are considered, IRR estimates are above 0.6 which indicates good reviewer agreement. Furthermore, we demonstrate that, with a small number of reviewers per proposal, zero estimates of IRR are possible even when the true value is not zero.

7.4.9 Model-based clustering with Missing Not At Random Data

Participants: Aude Sportisse.

Keywords: model-based clustering, generative models, missing data, medicine
Collaborations: Christophe Biernacki (Inria Lille), Claire Boyer (Sorbonne Université), Julie Josse (Inria Montpellier), Matthieu Marbac (Ensai Rennes)

With the increase of large datasets, the model-based clustering has become a very popular, flexible and interpretable methodology for data exploration in a well-defined statistical framework. However, in large scale data analysis, the problem of missing data is ubiquitous. We propose a novel approach by embedding missing data directly within model-based clustering algorithms. In particular, we consider the general case of Missing Not At Random (MNAR) values. We introduce in [44] a selection model for the joint distribution of data and missing-data indicator. It corresponds to a mixture model for the data distribution and a general MNAR model for the missing-data mechanism, for which the missingness may depend on the underlying classes (unknown) and/or the values of the missing variables themselves. A large set of meaningful MNAR sub-models is derived and the identifiability of the parameters is studied for each of the sub-models, which is usually a key issue for any MNAR proposals. The EM and Stochastic EM algorithms are considered for estimation. Finally, we perform empirical evaluations for the proposed sub-models on synthetic data (see e.g. Fig. 21) and we illustrate the relevance of our method on a medical register, the TraumaBase® dataset.

7.4.10 Unsupervized text clustering

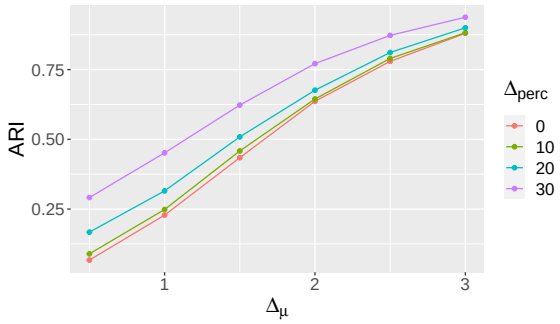


Figure 21: Relative effect on theoretical ARI of both the mixture component separation strength Δ_μ and the MNAR evidence Δ_{perc} ($\Delta_{\text{perc}} = 0$ is the MCAR case, the more $\Delta_{\text{perc}} = 0$ increases, the greater the deviation from the MCAR hypothesis.)

Participants: Michel Riveill, Xuchun Zhang.

Keywords: clustering, NLP, pharmacovigilance
Collaborations: Milou-Drici Daniel (Centre de Pharmacovigilance, Nice)

Detecting adverse drug events (ADEs) in the medical reports plays an pivot role in pharmacovigilance (PhV). Due to the shortage of well-trained professional to identify ADEs and the enormous amount of electoric health records (EHRs) produced each year, considering how expensive to annotate those EHRs for further study, we want to propose an unsupervised approach to distinguish the reports with and without ADEs descriptions.

We made the hypothesis that for any ADE, both the drug and the adverse effect are described within the same block of textual content. We defined henceforth "block" as the basic unit of textual content to analyse, which can be either whole document, paragraph, phrase, sentence, etc. Then we can define the problem as: For a given set of blocks B, suppose that we have had in advance the knowledge for the drug entities and symptom entities within each blocks (since drugs must have marketing authorisation and medical descriptions of symptoms have their universal codification), we want to separate the blocks with the description of ADE (noted as positive block b^+) from those who don't (noted as negative block b^-).

For the moment, the size of the "block » is limited to 500 tokens and consider for now the intra-sentence information. By the help of entity type information for drugs and symptoms in MADE[54] dataset, we tokenized only the context [49] around entities from the sentence. Then we fed the tokenized sentences to transformers model to encode each sentence into vectors representations. Finally A clustering algorithm is used for creating clusters of similar sentences. The ideal practice is to obtain a cluster with only positive blocks and another with only negative ones. Comparing to the supervised approach (Bag of words + Logistic Regression Classifier) with its f1-score as 0.8234 and f2-score as 0.8316, we found that both S-Bert [58] (with a f1-score of 0.6250 and f2-score of 0.6192) and BioBert [57] (f1-score as 0.7004 and f2 as 0.6955) can achieves relatively good results and latter even outperformed the former due to its domain specific knowledge.

7.4.11 Unsupervized text clustering

Participants: Michel Riveill, Mansour Zoubeirou A Mayaki.

Keywords: Deep Learning, Detection of weak signals, Incremental evolution of prediction models

Medicare fraud results in considerable losses for governments and insurance companies and results in higher premiums from clients. According to Insurance Europe, detected and undetected fraud costs

around 13 billion euros per year to European citizens [52]. In the field of healthcare insurance, in France the compulsory scheme detected over 261.2 million euros of fraudulent services in 2018, mainly due to healthcare professionals and healthcare establishments [50]. In the United States, according to the FBI, medicare fraud costs insurance companies between 21 billion and 71 billion US dollars per year [55]. In a context where reducing management costs is a real issue for healthcare insurers, the fight against fraud is a real expectation of the customers of professionals in the sector so that everyone receives a fair return for their contributions. This study aims to use artificial neural network based learners to detect fraudulent activities in medicare. Models based on statistical methods and machine learning make it possible to automatically build patterns and thus detect fraudulent activities effectively. The main difficulty in applying machine learning techniques in fraud detection or more generally anomaly detection is that you don't have enough data labeled as anomalous or fraudulent. In this study we use publicly available medicare data sets from the Centers for Medicare & Medicaid Services (CMS) for years 2017–2019 [51] and a medicare data available on kaggle [53]. The data sets contain three tables: hospitalization requests (Inpatient Data), outpatient care requests (Outpatient Data) and beneficiary information (Beneficiary Details Data). To detect medicare frauds, we propose to use multiple inputs deep neural networks based classifier with an autoencoder component. This architecture makes it possible to take into account many sources of data without mixing them and makes the classification task easier for the final model . Our results show that although baseline artificial neural network give good performances, they are outperformed by our multiple inputs neural networks [47]. We have shown that using an autoencoder to embed the provider level features gives better results and makes the classifiers more robust to class imbalance.

7.4.12 Continuous Latent Position Models for Instananeous Interactions

Participants: Marco Corneli.

Keywords: Latent Position Models, Dynamic Networks, Non-Homogeneous Poisson Process, Spatial Embeddings, Statistical Network Analysis
Collaborations: Riccardo Rastelli (UCD, Dublin)

In [39] we create a framework to analyze the timing and frequency of instantaneous interactions between pairs of entities. This type of interaction data is especially common nowadays, and easily available. Examples of instantaneous interactions include email networks, phone call networks and some common types of technological and transportation networks. Our framework relies on a novel extension of the latent position network model: we assume that the entities are embedded in a latent Euclidean space, and that they move along individual trajectories which are continuous over time. These trajectories are used to characterize the timing and frequency of the pairwise interactions. We discuss an inferential framework where we estimate the individual trajectories from the observed interaction data, and propose applications on artificial and real data. Figure 22 shows the evolving latent positions of a dynamic graph.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

8.1.1 Orange

Participants: Hugo Miralles, Michel Riveill
External participants: Tamara Tosic (Orange), Thierry Nagellen (Orange)
Value: 45000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Hugo Miralles on Distributed device-embedded classification and prediction in near-to-real time. In this thesis we study the problem of efficient classification and prediction of multivariate time-series captured by embedded devices by using joint data-model distributed algorithms for applications that preserve private data.

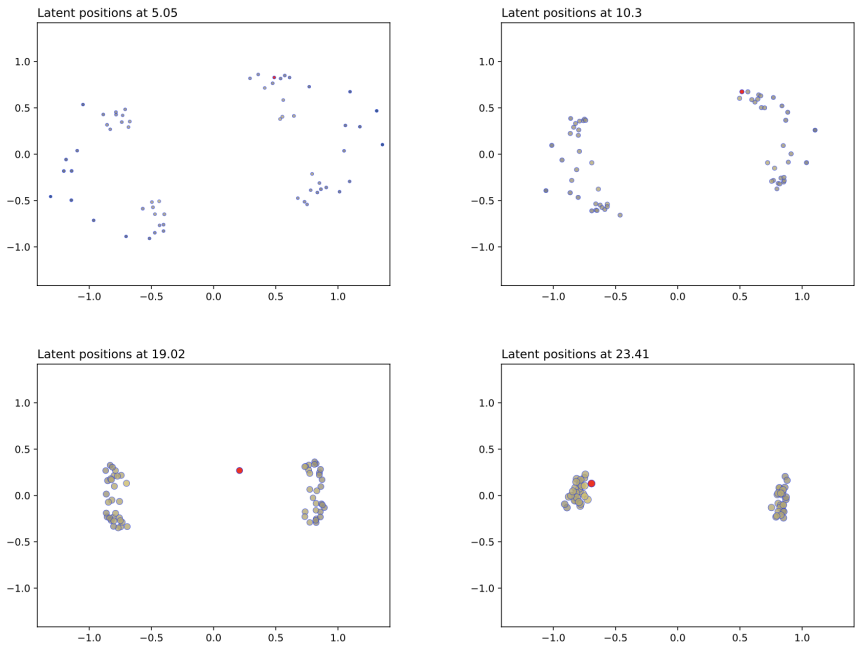


Figure 22: Snapashots of the evolving latent positions of a dynamic graph. Two communities emerge and a node (red) migrates between them.

8.1.2 NXP

Participants: Frederic Precioso, Charles Bouveyron, Baptiste Pouthier (Ph.D. candidate)
External participants: Laurent Pilati (NXP)
Value: 45000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Baptiste Pouthier on Deep Learning and Statistical Learning on audio-visual data for embedded systems.

8.1.3 Ezako

Participant: Frederic Precioso
External participants: Mireille Blay-Fornarino (Univ. Côte d'Azur), Yassine El Amraoui (Ezako - Univ. Côte d'Azur), Julien Muller (Ezako), Bora Kizil (Ezako)
Value: 45000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Yassine El Amraoui on Maximizing expert feedback in the detection of anomalies in time series.

8.1.4 Amadeus

Participant: Frederic Precioso
External participants: Nicolas Pasquier (Univ. Côte d'Azur), Tianshu Yang (Amadeus - Univ. Côte d'Azur), Antoine Hom (Amadeus), Laurent Dolle (Amadeus), Mickael Defoin-Platel (Amadeus), Luca Marchetti (Amadeus)
Value: 60000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Tianshu Yang on Semi-supervised clustering applied in revenue accounting.

8.1.5 Detection and characterization of salient moments for automatic summaries

Participants: Melissa Sanabria, Frédéric Precioso.

External Participants: Thomas Menguy (Wildmoka)
Value: 45000 EUR
Keywords: Video Summarization, Multimodal data, Soccer match data.

We have designed an architecture, introducing a Multiple Instance Learning method that takes into account the sequential dependency among events and a hierarchical multimodal attention layer that grasps the importance for each event in an action. We evaluate our approach on matches from two professional European soccer leagues, showing its capability to identify the best actions for automatic summarization by comparing with real summaries made by human operators.

8.1.6 Naval Group

Participants: Pierre-Alexandre Mattei
External Participants: Alexandre Gensse, Quentin Oliveau (Naval Group)
Value: 125000 EUR
Keywords: Semi-supervised learning

The goal of this project will be the development of an open-source Python library for semi-supervised learning, via the hiring of a research engineer.

8.2 Bilateral grants with industry

8.2.1 Grant from the Novo Nordisk foundation

Participant: Pierre-Alexandre Mattei
External participants: Wouter Boomsma (University of Copenhagen), Jes Frellsen (Technical University of Denmark), Søren Hauberg (principal investigator, Technical University of Denmark), and Ole Winther (Technical University of Denmark)
Value: 180000 DKK (≈ 24000 EUR)

The object of this grant was the organisation of the 2nd Copenhagen Workshop on Generative Models and Uncertainty Quantification (GenU), that was held in October 2021 in Copenhagen.

9 Partnerships and cooperations

9.1 International initiatives

Informal international partners The Maasai team has informal relationships with the following international teams:

- Department of Statistics of the University of Washington, Seattle (USA) through collaborations with Elena Erosheva and Adrian Raftery,
- SAILAB team at Universita di Sienna, Sienna (Italy) through collaborations with Marco Gori,
- School of Mathematics and Statistics, University College Dublin (Ireland) through the collaborations with Brendan Murphy, Riccardo Rastelli and Michael Fop,
- Department of Computer Science, University of Tübingen (Germany) through the collaboration with Ulrike von Luxburg,
- Université Laval, Québec (Canada) through the Research Program DEEL (DEpendable and Explainable Learning) with François Laviolette and Christian Gagné, and through a FFCR funding with Arnaud Droit (including the planned supervision of two PhD students in 2022),
- DTU Compute, Technical University of Denmark, Copenhagen (Denmark), through collaborations with Jes Frellsen and his team (including the co-supervision of a PhD student in Denmark: Hugo Sénétaire).

9.1.1 Participation in other International Programs

DEpendable Explainable Learning Program (DEEL), Québec, Canada

Participants: Frederic Precioso

Collaborations: François Laviolette (Prof. U. Laval), Christian Gagné (Prof. U. Laval)

The DEEL Project involves academic and industrial partners in the development of dependable, robust, explainable and certifiable artificial intelligence technological bricks applied to critical systems. We are involved in the Workpackage Robustness and the Workpackage Interpretability, in the co-supervision of several PhD thesis, Post-docs, and Master internships.

CHU Québec–Laval University Research Centre, Québec, Canada

Participants: Pierre-Alexandre Mattei, Frederic Precioso, Louis Ohl (doctorant)

Collaborations: Arnaud Droit (Prof. U. Laval), Mickael Leclercq (Chercheur U. Laval), Khawla Seddiki (doctorante, U. Laval)

This collaboration framework covers several research projects: one project is related to the PhD thesis of Khawla Seddiki who works on Machine Learning/Deep Learning methods for classification and analysis of mass spectrometry data; another project is related to the France Canada Research Fund (FCRF) which provides the PhD funding of Louis Ohl, co-supervised by all the collaborators. This project investigates Machine Learning solutions for Aortic Stenosis (AS) diagnosis.

SAILAB: Lifelong learning in computer vision Participants: Lucile Sassatelli and Frédéric Precioso (UCA)

Keywords: computer vision, lifelong learning, focus of attention in vision, virtual video environments.

Collaborations: Dario (Universität Erlangen-Nürnberg), Alessandro Betti (UNISI), Stefano Melacci (UNISI), Matteo Tiezzi (UNISI), Enrico Meloni (UNISI), Simone Marullo (UNISI).

This collaboration concerns the current hot machine learning topics of Lifelong Learning, "on developing versatile systems that accumulate and refine their knowledge over time"), or continuous learning which targets tackling catastrophic forgetting via model adaptation. The most important expectations of this research is that of achieving object recognition visual skills by a little supervision, thus overcoming the need for the expensive accumulation of huge labelled image databases.

9.2 European initiatives

9.2.1 FP7 & H2020 Projects

Maasai is one of the 3IA-UCA research teams of **AI4Media**, one of the 4 ICT-48 Center of Excellence in Artificial Intelligence which has started in September 2020. There are 30 partners (Universities and companies), and 3IA-UCA received about 325k€.

9.3 National initiatives

Institut 3IA Côte d'Azur Following the call of President Macron to found several national institutes in AI, we presented in front of a international jury our project for the Institut 3IA Côte d'Azur in April 2019. The project was selected for funding (50 M€ for the first 4 years, including 16 M€ from the PIA program) and started in september 2019. Charles Bouveyron and Marco Gori are two of the 29 3IA chairs which were selected *ab initio* by the international jury and Pierre-Alexandre Mattei was awarded a 3IA chair in 2021. Charles Bouveyron is also the Director of the institute since January 2021, after being the Deputy Scientific Director on 2019-2020. The research of the institute is organized around 4 thematic axes: Core elements of AI, Computational Medicine, AI for Biology and Smart territories. The Maasai reserch team is totally aligned with the first axis of the Institut 3IA Côte d'Azur and also contributes to the 3 other axes through applied collaborations. The team has 7 Ph.D. students and postdocs who are directly funded by the institute.

Web site: 3ia.univ-cotedazur.eu

9.4 Regional initiatives

Parc National du Mercantour

Participants: Charles Bouveyron, Frédéric Precioso and Fanny Simoẽs

Keywords: Deep learning, image recognition,

Collaborators: Nathalie Siefert and Stéphane Combeau (Parc National du Mercantour)

The team started in 2021 a collaboration with the Parc National du Mercantour to exploit the camera-traps installed in the Park to monitor and conserve wild species. We developed, in collaboration with the engineer team of Institut 3IA Côte d'Azur, an AI pipeline allowing to automically detect, callsify and count specific endangered wild species in camera-trap videos. A demonstrator of the methodology has been presented to the general public at *Le Fête des Sciences* in Antibes in October 2021.

Centre de pharmacovigilance, CHU Nice

Participants: Charles Bouveyron, Marco Corneli, Giulia Marchello, Michel Riveill, Xuchun Zhang

Keywords: Pharmacovigilance, co-clustering, count data, text data

Collaborateurs: Milou-Daniel Drici, Audrey Freysse, Fanny Serena Romani

The team works very closely with the Regional Pharmacovigilance Center of the University Hospital Center of Nice (CHU) through several projects. The first project concerns the construction of a dashboard to classify spontaneous patient and professional reports, but above all to report temporal breaks. To this end, we are studying the use of dynamic co-classification techniques to both detect significant ADR patterns and identify temporal breaks in the dynamics of the phenomenon. The second project focuses on the analysis of medical reports in order to extract, when present, the adverse events for characterization. After studying a supervised approach, we are studying techniques requiring fewer annotations.

Interpretability for automated decision services

Participants: Damien Garreau, Frédéric Precioso

Keywords: interpretability, deep learning

Collaborators: Greger Ottosson (IBM)

Businesses rely more and more frequently on machine learning to make automated decisions. In addition to the complexity of these models, a decision is rarely by using only one model. Instead, the crude reality of business decision services is that of a jungle of models, each predicting key quantities for the problem at hand, that are then agglomerated to produce the final decision, for instance by a decision tree. In collaboration with IBM, we want to provide principled methods to obtain interpretability of these automated decision processes.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

GenU workshop Pierre-Alexandre Mattei co-organised the second edition of the Generative Models and Uncertainty Quantification (GenU) workshop. This small-scale workshop was held physically in Copenhagen (October 12-13).
Web site: genu.ai/2021/

52ème Journées de Statistique de la SFDS Charles Bouveyron, Marco Corneli, Damien Garreau, Pierre-Alexandre Mattei and Frédéric Precioso are members of the organization comittee of the conference "52ème Journées de Statistique", which is the annual conference of the French Statistical Association (SFdS). This conference usually gathers more than 400 statisticians from academic and industry. The 2020 edition was initially planned in Nice on 25-29 May 2020. Due to the pandemic, the conference has been postponed by one year and has been held in Nice on 7-11 June 2021. Charles Bouveyron was the President of the organization comittee.
Web site: jds2021.sciencesconf.org

10.1.2 Scientific events: selection

Member of the conference program committees

- Frédéric Precioso is a member of the program committee of the the conference "52ème Journées de Statistique", which is the annual conference of the French Statistical Association (SFdS). The

2020 edition was initially planned in Nice on 25-29 May 2020. Due to the pandemic, the conference has been postponed by one year and has been held in Nice on 7-11 June 2021.

Member of the editorial boards

- Charles Bouveyron is Associate Editor for the Annals of Applied Statistics since 2016.

Reviewer - reviewing activities All permanent members of the team are serving as reviewers for the most important journals and conferences in statistical and machine learning, such as (non exhaustive list):

- International journals:
  - Annals of Applied Statistics,
  - Statistics and Computing,
  - Journal of the Royal Statistical Society, Series C,
  - Journal of Computational and Graphical Statistics,
  - Journal of Machine Learning Research
- International conferences:
  - Neural Information Processing Systems (Neurips),
  - International Conference on Machine Learning (ICML),
  - International Conference on Learning Representations (ICLR),
  - International Joint Conference on Artificial Intelligence (IJCAI),
  - International Conference on Artificial Intelligence and Statistics (AISTATS),
  - International Conference on Computer Vision and Pattern Recognition

10.1.3 Invited talks

- Charles Bouveyron was invited for a keynote talk at the 1st French-German Machine Learning Symposium.
- Pierre-Alexandre Mattei was invited to give a talk at the technical University of Denmark in October 2021.

10.1.4 Leadership within the scientific community

- Charles Bouveyron is the Director of the Institut 3IA Côte d'Azur since January 2021.

10.1.5 Scientific expertise

Most permanent members of the team are serving as experts for the ANR or foreign research agencies.

10.1.6 Research administration

- Frédéric Precioso is the Scientific Responsible for AI at the French Research Agency (ANR) since September 2019.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

C. Bouveyron, D. Garreau, F. Precioso and M. Riveill are professors at Université Côte d'Azur and therefore handle usual teaching duties. M. Corneli and P.-A. Mattei are also teaching around 50h per year at Université Côte d'Azur. P.-A. Mattei is also teaching a graphical models course at the MVA masters from ENS Paris Saclay.

C. Bouveyron (up to august 2020) and M. Riveill (since September 2020) are responsible for the MSc. Data Sciences and Artificial Intelligence at Université Côte d'Azur.

10.2.2 Supervision

PhD students, postdocs, and interns of the team are listed in Section 1. Additionally, members of the team supervise several Masters projects, in particular from the MSc. Data Sciences and Artificial Intelligence at Université Côte d'Azur.

10.3 Popularization

10.3.1 Interventions

- Frederic Precioso, Charles Bouveyron, Fanny Simoes and Jonathan Torres Sanchez have developed a demonstrator for general public on the recognition and monitoring of wild species in the French National Park of Mercantour. This demonstrator has been exhibited during the "Fête des Sciences" in Antibes - Juan-les-Pins on October 2021.
- Frederic Precioso has developed an experimental platform both for research projects and scientific mediation on the topic of autonomous cars. This platform is currently installed in the "Maison de l'Intelligence Artificielle" where high school students have already experimented coding autonomous remote control cars (https://maison-intelligence-artificielle.com/experimenter-projets-ia/).
- Charles Bouveyron, Fanny Simoes and Silvia Bottini have developed an interactive software allowing to visualise the relationships between pollution and a health disease (dispnea) in the Région Sud. This platform is currently installed in the "Maison de l'Intelligence Artificielle".

11 Scientific production

11.1 Major publications

[1] C. Bouveyron, J. Jacques, A. Schmutz, F. Simoes and S. Bottini. 'Co-Clustering of Multivariate Functional Data for the Analysis of Air Pollution in the South of France'. In: *Annals of Applied Statistics* (2021). URL: https://hal.archives-ouvertes.fr/hal-02862177.

[2] E. A. Erosheva, P. Martinková and C. J. Lee. 'When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review'. In: *Journal of the Royal Statistical Society: Series A Statistics in Society* 184 (20th Apr. 2021), pp. 904–919. DOI: 10.1111/rssa.12681. URL: https://hal.archives-ouvertes.fr/hal-03522263.

[3] D. Garreau and D. Mardaoui. 'What does LIME really see in images?' In: ICML 2021 - 38th International Conference on Machine Learning. Virtual Conference, United States, 18th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03233014.

[4] N. B. Ipsen, P.-A. Mattei and J. Frellsen. 'not-MIWAE: Deep Generative Modelling with Missing not at Random Data'. In: ICLR 2021 - International Conference on Learning Representations. Virtual, Austria, 2021. URL: https://hal.inria.fr/hal-03044124.

[5] N. Jouvin, C. Bouveyron and P. Latouche. 'A Bayesian Fisher-EM algorithm for discriminative Gaussian subspace clustering'. In: *Statistics and Computing* (23rd May 2021). DOI: 10.1007/s11222-021-10018-6. URL: https://hal.archives-ouvertes.fr/hal-03047930.

[6] D. Mardaoui and D. Garreau. 'An Analysis of LIME for Text Data'. In: *AISTATS 2021 - 24th International Conference on Artificial Intelligence and Statistics*. AISTATS 2021 - 24th International Conference on Artificial Intelligence and Statistics. Vienne, Austria, 13th Apr. 2021. URL: https://hal.archives-ouvertes.fr/hal-02935171.

[7] M. F. Romero Rondon, L. Sassatelli, R. Aparicio-Pardo and F. Precioso. 'TRACK: A New Method from a Re-examination of Deep Architectures for Head Motion Prediction in 360-degree Videos'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1. DOI: 10.1109/TPAMI.2021.3070520. URL: https://hal.archives-ouvertes.fr/hal-03193067.

[8] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, M. Wisztorski, M. Salzet, I. Fournier and A. Droit. 'Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification'. In: *Nature Communications* 11.1 (Dec. 2020). DOI: 10.1038/s41467-020-19354-z. URL: https://hal.archives-ouvertes.fr/hal-03132326.

[9] M. Tiezzi, S. Melacci, A. Betti, M. Maggini and M. Gori. 'Focus of Attention Improves Information Transfer in Visual Features'. In: NeurIPS 2020 - 34th Conference on Neural Information Processing Systems. Vancouver / Online, Canada, 6th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-02878372.

[10] C. Vincent-Cuaz, T. Vayer, R. Flamary, M. Corneli and N. Courty. 'Online Graph Dictionary Learning'. In: ICML 2021 - 38th International Conference on Machine Learning. ICML 2021 - 38th International Conference on Machine Learning. Virtual Conference, United States, 2021. URL: https://hal.archives-ouvertes.fr/hal-03140349.

11.2 Publications of the year

International journals

[11] S. Allassonnière and J. Chevallier. 'A New Class of Stochastic EM Algorithms. Escaping Local Maxima and Handling Intractable Sampling'. In: *Computational Statistics and Data Analysis* 159 (July 2021), p. 107159. DOI: 10.1016/j.csda.2020.107159. URL: https://hal.archives-ouvertes.fr/hal-02044722.

[12] A. Bobasheva, F. Gandon and F. Precioso. 'Learning and Reasoning for Cultural Metadata Quality'. In: *Journal on Computing and Cultural Heritage* (2022). URL: https://hal.archives-ouvertes.fr/hal-03363442.

[13] C. Bouveyron, A. Casa, E. Erosheva and G. Menardi. 'Co-clustering of Time-Dependent Data via the Shape Invariant Model'. In: *Journal of Classification* (2021). DOI: 10.1007/s00357-021-09402-8. URL: https://hal.archives-ouvertes.fr/hal-03370436.

[14] C. Bouveyron, J. Jacques, A. Schmutz, F. Simoes and S. Bottini. 'Co-Clustering of Multivariate Functional Data for the Analysis of Air Pollution in the South of France'. In: *Annals of Applied Statistics* (2021). URL: https://hal.archives-ouvertes.fr/hal-02862177.

[15] J. Chevallier, V. Debavelaere and S. Allassonnière. 'A coherent framework for learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data'. In: *SIAM Journal on Imaging Sciences* 14.1 (2021), pp. 349–388. DOI: 10.1137/20M1328026. URL: https://hal.archives-ouvertes.fr/hal-01646298.

[16] E. Côme, P. Latouche, N. Jouvin and C. Bouveyron. 'Hierarchical clustering with discrete latent variable models and the integrated classification likelihood'. In: *Advances in Data Analysis and Classification* (2021). DOI: 10.1007/s11634-021-00440-z. URL: https://hal.archives-ouvertes.fr/hal-02530705.

[17] E. A. Erosheva, P. Martinková and C. J. Lee. 'When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review'. In: *Journal of the Royal Statistical Society: Series A Statistics in Society* 184 (20th Apr. 2021), pp. 904–919. DOI: 10.1111/rssa.12681. URL: https://hal.archives-ouvertes.fr/hal-03522263.

[18] M. Fop, P.-A. Mattei, C. Bouveyron and T. B. Murphy. 'Unobserved classes and extra variables in high-dimensional discriminant analysis'. In: *Advances in Data Analysis and Classification* (2021). URL: https://hal.archives-ouvertes.fr/hal-03132362.

[19] D. Fraix-Burnet, C. Bouveyron and J. Moultaka. 'Unsupervised classification of SDSS galaxy spectra'. In: *Astronomy and Astrophysics - A&A* 649 (2021), A53. DOI: 10.1051/0004-6361/202040046. URL: https://hal.archives-ouvertes.fr/hal-03162811.

[20] N. Jouvin, C. Bouveyron and P. Latouche. 'A Bayesian Fisher-EM algorithm for discriminative Gaussian subspace clustering'. In: *Statistics and Computing* (23rd May 2021). DOI: 10.1007/s11222-021-10018-6. URL: https://hal.archives-ouvertes.fr/hal-03047930.

[21] A. Koukounari, H. Jamil, E. Erosheva, C. Shiff and I. Moustaki. 'Latent Class Analysis: Insights about design and analysis of schistosomiasis diagnostic studies'. In: *PLoS Neglected Tropical Diseases* 15 (4th Feb. 2021). DOI: 10.1371/journal.pntd.0009042. URL: https://hal.archives-ouvertes.fr/hal-03522272.

[22] D. Liang, M. Corneli, C. Bouveyron and P. Latouche. 'DeepLTRS: A deep latent recommender system based on user ratings and reviews'. In: *Pattern Recognition Letters* (Dec. 2021). DOI: 10.1016/j.patrec.2021.10.022. URL: https://hal.archives-ouvertes.fr/hal-03021362.

[23] L. Rendsburg and D. Garreau. 'Comparison-based centrality measures'. In: *International Journal of Data Science and Analytics* 11 (12th Apr. 2021), pp. 243–259. DOI: 10.1007/s41060-021-00254-4. URL: https://hal.archives-ouvertes.fr/hal-03233015.

[24] M. F. Romero Rondon, L. Sassatelli, R. Aparicio-Pardo and F. Precioso. 'TRACK: A New Method from a Re-examination of Deep Architectures for Head Motion Prediction in 360-degree Videos'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1. DOI: 10.1109/TPAMI.2021.3070520. URL: https://hal.archives-ouvertes.fr/hal-03193067.

[25] T. Yang, N. Pasquier and F. Precioso. 'Semi-supervised Consensus Clustering Based on Closed Patterns'. In: *Knowledge-Based Systems*. Article 107599 235 (10th Jan. 2022). DOI: 10.1016/j.knosys.2021.107599. URL: https://hal.archives-ouvertes.fr/hal-03402517.

International peer-reviewed conferences

[26] D. Garreau and D. Mardaoui. 'What does LIME really see in images?' In: ICML 2021 - 38th International Conference on Machine Learning. Virtual Conference, United States, 18th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03233014.

[27] N. B. Ipsen, P.-A. Mattei and J. Frellsen. 'not-MIWAE: Deep Generative Modelling with Missing not at Random Data'. In: ICLR 2021 - International Conference on Learning Representations. Virtual, Austria, 2021. URL: https://hal.inria.fr/hal-03044124.

[28] D. Mardaoui and D. Garreau. 'An Analysis of LIME for Text Data'. In: *AISTATS 2021 - 24th International Conference on Artificial Intelligence and Statistics*. AISTATS 2021 - 24th International Conference on Artificial Intelligence and Statistics. Vienne, Austria, 13th Apr. 2021. URL: https://hal.archives-ouvertes.fr/hal-02935171.

[29] E. Meloni, L. Pasqualini, M. Tiezzi, M. Gori and S. Melacci. 'SAILenv: Learning in Virtual Visual Environments Made Simple'. In: ICPR 2020 - 25th International Conference on Pattern Recognition. Milan / Virtual, Italy, 10th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-02965715.

[30] B. Pouthier, L. Pilati, L. Gudupudi, C. Bouveyron and F. Precioso. 'Active Speaker Detection as a Multi-Objective Optimization with Uncertainty-Based Multimodal Fusion'. In: Interspeech 2021. Brno, Czech Republic: ISCA, 30th Aug. 2021, pp. 2381–2385. DOI: 10.21437/Interspeech.2021-80. URL: https://hal.archives-ouvertes.fr/hal-03345281.

[31] M. Sanabria, F. Precioso and T. Menguy. 'Hierarchical Multimodal Attention for Deep Video Summarization'. In: ICPR 2020 - 25th International Conference on Pattern Recognition. Milan / Virtual, Italy, 10th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-02964209.

[32] C. Vincent-Cuaz, T. Vayer, R. Flamary, M. Corneli and N. Courty. 'Online Graph Dictionary Learning'. In: ICML 2021 - 38th International Conference on Machine Learning. ICML 2021 - 38th International Conference on Machine Learning. Virtual Conference, United States, 2021. URL: https://hal.archives-ouvertes.fr/hal-03140349.

Conferences without proceedings

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

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

[35] H. Miralles, T. Tosic and M. Riveill. 'Communication-efficient Federated Learning through Clustering optimization'. In: SophI.A. Summit. Biot, France, 17th Nov. 2021. URL: https://hal.inria.fr/hal-03479640.

Scientific book chapters

[36] L. Vanni and F. Precioso. 'Deep learning et description des textesArchitecture méthodologique'. In: *L'intelligence artificielle des textes*. Champion, 2021, pp. 15–72. URL: https://hal.archives-ouvertes.fr/hal-03230830.

Reports & preprints

[37] S. Bartels, W. Boomsma, J. Frellsen and D. Garreau. *Kernel-Matrix Determinant Estimates from stopped Cholesky Decomposition*. 30th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03359274.

[38] G. Ciravegna, F. Precioso and M. Gori. *Knowledge-driven Active Learning*. 15th Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03527128.

[39] M. Corneli and R. Rastelli. *Continuous Latent Position Models for Instantaneous Interactions*. 7th Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03517392.

[40] Y. Gu, E. A. Erosheva, G. Xu and D. B. Dunson. *Dimension-Grouped Mixed Membership Models for Multivariate Categorical Data*. 12th Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03522275.

[41] R. Khouja, P.-A. Mattei and B. Mourrain. *Tensor decomposition for learning Gaussian mixtures from moments*. 1st June 2021. URL: https://hal.archives-ouvertes.fr/hal-03244448.

[42] G. Lopardo, D. Garreau, F. Precioso and G. Ottosson. *SMACE: A New Method for the Interpretability of Composite Decision Systems*. 15th Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-03527129.

[43] G. Marchello, A. Fresse, M. Corneli and C. Bouveyron. *Co-clustering of evolving count matrices in pharmacovigilance with the dynamic latent block model*. 19th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03146769.

[44] A. Sportisse, C. Biernacki, C. Boyer, J. Josse, M. Marbac Lourdelle, G. Celeux and F. Laporte. *Model-based Clustering with Missing Not At Random Data*. 19th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03494674.

[45] L. Vanni, M. Corneli, D. Mayaffre and F. Precioso. *From text saliency to linguistic objects: learning linguistic interpretable markers with a multi-channels convolutional architecture*. 15th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03142170.

[46] C. Vincent-Cuaz, R. Flamary, M. Corneli, T. Vayer and N. Courty. *Semi-relaxed Gromov Wasserstein divergence with applications on graphs*. 13th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03376923.

[47] M. ZOUBEIROU A MAYAKI and M. Riveill. *Multiples inputs neural nets for Medicare fraud detection*. 22nd Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03500411.

Other scientific publications

[48] X. Zhang and M. Riveill. 'Unsupervised Adverse Drug Event related document detection with Bert-based model'. In: Sophia Summit. Sophia Antipolis, France, 17th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03519982.

11.3 Cited publications

[49] E. Agichtein and L. Gravano. 'Snowball: extracting relations from large plain-text collections'. In: *Proceedings of the fifth ACM conference on Digital libraries - DL '00*. the fifth ACM conference. San Antonio, Texas, United States: ACM Press, 2000, pp. 85–94. DOI: 10.1145/336597.336644. URL: http://portal.acm.org/citation.cfm?doid=336597.336644 (visited on 11/06/2021).

[50] *Bilan 2018 des actions de lutte contre la fraude et actions de contrôles*. Caisse nationale de l'Assurance Maladie. URL: https://www.ameli.fr/sites/default/files/2019-10-01-dp-controles-fraudes.pdf. (accessed: 10.05.2021).

[51] *Centers For Medicare & Medicaid Services. Medicare fee-for-service provider utilization & payment data physician and other supplier public use file: a methodological overview*. Centers For Medicare & Medicaid Services. 2018. URL: https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data/physician-and-other-supplier. (accessed: 01.11.2020).

[52] *Fraud prevention*. insurance europe. URL: https://www.insuranceeurope.eu/priorities/23/fraud-prevention. (accessed: 10.08.2021).

[53] *HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS*. Kaggle. URL: https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis/metadata. (accessed: 10.05.2021).

[54] A. Jagannatha, F. Liu, W. Liu and H. Yu. 'Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0)'. In: *Drug Safety* 42.1 (2019), pp. 99–111. DOI: 10.1007/s40264-018-0762-z.

[55] J. M. Johnson and T. M. Khoshgoftaar. 'Medicare fraud detection using neural networks'. In: *Journal of Big Data* 6.1 (2019), pp. 1–35.

[56] Y. LeCun, Y. Bengio and G. Hinton. 'Deep learning'. In: *nature* 521.7553 (2015), pp. 436–444.

[57] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang. 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining'. In: *Bioinformatics* (Sept. 2019). arXiv: 1901.08746, btz682. DOI: 10.1093/bioinformatics/btz682.

[58] N. Reimers and I. Gurevych. 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'. In: *arXiv:1908.10084 [cs]* (27th Aug. 2019). arXiv: 1908.10084. URL: http://arxiv.org/abs/1908.10084 (visited on 08/10/2020).