Statistical learning on networks and more

Manuscript for the obtention of the
Habilitation à diriger des recherches

Tabea Rebafka

Defended on June 22, 2023.
Jury composed of:

Christophe Ambroise	Université d'Évry, PR	Rapporteur
Florence Forbes	INRIA Grenoble-Alpes, DR	Examinatrice
Eric Kolaczyk	McGill University, PR	Rapporteur
Brendan Murphy	University College Dublin, PR	Rapporteur
Stéphane Robin	Sorbonne Université, PR	Président du jury
Nathalie Vialaneix	INRAE Toulouse, DR	Examinatrice

Sorbonne Université
Paris, France
June 2023

Ackowledgment

First and foremost, I would like to extend my warmest thanks to the three referees for taking the time to read and comment on this manuscript. I would also like to thank all the members of the committee for their interest in my work. It is a great honour that you have agreed to participate in my jury, and as you are all very busy and have many important commitments, I really appreciate your availability.

I wrote the manuscript during my délégation at MaIAGE, INRAE, where I was warmly welcomed and allowed to take time for this work.

The habilitation is the fruit of several years of research, mostly carried out at Sorbonne Université, at the LPMA from 2010 to 2017, then at the LPSM. I am very grateful for the excellent working conditions and wonderful colleagues that have enabled me to develop professionally. Most of my achievements have been made possible thanks to others who have supported me or opened up new opportunities. It's hard to list all of them without forgetting someone. I'll mention just a few: Catherine Matias introduced me to the field of statistical network analysis; Nataliya Sokolovska is a never-ending source of ideas and new projects; Arnaud Guyader entrusted me with the direction of the Master in Statistics; Lorenzo Zambotti appointed me deputy director of the LPSM; Estelle Kuhn invited me at INRAE for a *délégation*; Ariane Marandon, Arsen Sultanov and Sara Rejeb accepted the adventure of doing a Ph.D. under my supervision in cooperation with co-supervisers Catherine Duveau, Etienne Roquain, Jean-Claude Crivello and Nataliya Sokolovska.

I am also very grateful to my co-authors. I learned a lot from all of them and we often spent a great time together: Antoine Souloumiac, Ariane Marandon, Arsen Sultanov, Catherine Duveau, Catherine Matias, Céline Lévy-Leduc, Estelle Kuhn, Etienne Roquain, Fanny Villers, Ismaël Castillo, Jean-Claude Crivello, Fabienne Comte, François Roueff, Malika Kharouf, Maurice Charbit, Max Feinberg, Nataliya Sokolovska, Sara Rejeb, Stephan Clémençon.

My greatest thanks naturally go to my friends and family, which stand by me whatever happens and whatever I do. Finally, I give a big hug to my lovely kids William and Raphaël.

Contents

Chapter 1

Introduction

This manuscript gives an overview of my research since my doctoral thesis and it is written in view of the obtention of the *Habilitation à diriger des recherches*. It is a presentation of the main ideas of my work and of the importance of the obtained results with respect to the existing literature. The manuscript is not meant to be exhaustive. For technical details and a more rigorous presentation of the methods and the results we refer the reader to the published articles.

Generally speaking, my research is driven by practical applications and by problems that motivate me to seek concrete mathematical solutions. In general, one important aspect of my work is modelling. A model is "good" when it reflects reality, is relatively simple, interpretable and as flexible as possible to cover a wide range of cases. The other important part of my work is the development of statistical methods as inference algorithms and testing procedures. It is clear that algorithms should be fast and reliable, but this is challenging to achieve when models are complex, or when available data sets become huge. In short, I enjoy developing new statistical models, that have a practical use, as well as efficient methods and algorithms, that have a theoretical foundation if possible.

Concerning my research fields, my Ph.D. was concerned with latent variable models and inverse problems for a specific problem in physics. Later, I worked on nonparametric estimation problems with automatic model selection as well as on statistical learning problems in signal processing such as regularization methods and dimension reduction. Then, I discovered random graph models, in which I am still interested. For network data, I study questions related to modelling and algorithmic development. Globally, my research activity after my doctoral thesis can be organized into three fields, which correspond to Chapters 2, 3 and 4 of this manuscript: statistical learning on networks, nonparametric density estimation in inverse problems and various topics in machine learning.

My research results are published in journals and conferences, and most of my code is publicly available. My very first research paper [RCF07] dates back to my Master thesis that I did at INRA in 2006. The papers [RRS10, RRS11] and the patent [RRS09] cover my Ph.D. at Télécom ParisTech and CEA Saclay from 2006 to 2009. The papers [RLLC11a, RLLC11b] correspond to my Postdoc at Télécom ParisTech the year after my doctoral thesis. And all the other work

[CR12, RR15, CR16, CR17, MRV18, KRS18, KMR20, RRV22, RDR22, MRRS22, Reb22] was done during my stay at Sorbonne University, formerly Université Pierre et Marie Curie, since 2010.

This chapter provides a resumé of my scientific accomplishments and publications of the last years, which are presented in more detail in the rest of the manuscript.

1.1 Statistical learning on networks

All my contributions to the field of statistical network analysis are related to the popular stochastic block model (SBM). Defining a very rich family of probability distributions, this random graph model accommodates most heterogeneous network topologies encountered in practice. Moreover, model parameters are highly interpretable, and as such of much interest for applications.

In a first work, which is detailed in Section 2.1, we propose the very first time-continuous extension of the SBM for dynamic temporal networks by introducing inhomogeneous Poisson processes with intensities depending on the latent block structure of the SBM. For the inference, a semiparametric variational Expectation-Maximization algorithm is proposed including a nonparametric M-step in form of adaptive estimators of the Poisson process intensities.

Section 2.2 describes how to use mini-batch sampling to speed up the Monte Carlo Markov Chain Stochastic Approximation Expectation Maximization (MCMC-SAEM) algorithm for inference in the SBM and in other general latent variable models. Theoretical results on the convergence of the proposed procedure are provided. We also illustrate that the choice of the mini-batch size can be optimized under the constraint of a limited computing time budget.

In Section 2.3 we use a random graph model to build a powerful multiple testing procedure for paired null hypotheses, when tests have to be performed for all pairs of entities in a population. The true/false null constellation is assumed to be structured according to an unobserved graph and we improve the power of the testing procedure by learning the graph structure modelled by a SBM. The procedure is shown to be nearly optimal and the results hold in the finite-sample setting, which is a novelty in the domain.

Section 2.4 is devoted to the clustering of a set of networks. A model-based clustering approach based on a finite mixture of stochastic block models is proposed together with an efficiently implemented hierarchical agglomerative algorithm. The algorithm is shown to outperform classical distance-based graph clustering methods. As a byproduct of the hierarchical algorithm, we propose a new tool to match node labels of two stochastic block models to overcome the label-switching problem of the SBM.

1.2 Nonparametric density estimation in inverse problems

Part of my research is devoted to density estimation in specific inverse problems and this work is presented in Chapter 3. In Section 3.1, we consider continuous mixture models and the problem of estimating the mixing density. We propose an orthogonal series estimator based on

polynomial approximations and show that in case of an exponential mixture it is optimal in the sense that it achieves the minimax rate over a specific collection of smoothness classes.

Motivated by fluorescence lifetime measurements, we investigate density estimation in statistical models characterized by some nonlinear distortion. An adaptive nonparametric estimator for the so-called pile-up model is proposed and oracle-type risk bounds for the mean integrated squared error are provided. In a follow-up work, we propose and compare several kernel and projection estimation strategies for biased data models, that come with different data-driven model and bandwidth selection methods. It is shown that the estimators perform an automatic finite-sample bias-variance tradeoff and a numerical study provides a comparison of the estimators and reveals an interesting gap between theory and practice. This work is presented in Section 3.2.

1.3 Topics in machine learning

There are various topics in machine learning and signal processing to which I have contributed. First, to recover the waveform of an intercepted radar signal and estimate its direction of arrival, we propose a regularized orthogonal matching pursuit algorithm with structured sparsity patterns. The method is suitable for very low signal-to-noise ratios (Section 4.1).

Second, building on results from random matrix theory, a novel estimator of the dimension of the subspace, where an observed noisy signal lives in, is proposed and shown to be consistent. Applications to biomedical data sets illustrate the improvements compared to the state of the art (Section 4.2).

Third, for self-organizing maps we propose an extension that handles missing data. The approach is based on a new objective function so that missing values are learned at the same time as the map. The method is used to analyze production data of aircraft engines (Section 4.3).

Fourth, when clustering individuals, misclassifications may be very costly in practical applications. We propose a model-based clustering approach with an abstention option such that only a part of the sample is clustered. The goal is to keep the misclustering rate below some nominal level. Our procedure, developed in the finite mixture model framework, is shown to have nearly optimal power and bootstrap even improves the performance (Section 4.4).

Chapter 2

Statistical learning on networks

Statistical network analysis is an active field of research, with increasing importance over the last years due to the emergence of network-structured data in a large variety of fields of application [1, 2]. Networks provide powerful descriptions of relations and interactions among a given set of entities, but they are complex mathematical objects, which are hard to analyze due to their involved dependance structures. In this chapter a presentation of my four main contributions to the field is provided. They are all related to the popular stochastic block model (SBM) which is an easily interpretable model that accommodates most heterogeneous networks observed in practice. We start by recalling the definition of this model and fixing notations.

Consider a network with n vertices. The *stochastic block model* (SBM) can be defined for both directed and undirected interactions. Denote $\mathcal{R}$ the set of all dyads in the network, that is, if the graph is directed, $\mathcal{R} = \{(i,j) : i \neq j\} \subset [\![n]\!]^2$, and $\mathcal{R} = \{(i,j) : i < j\} \subset [\![n]\!]^2$ if it is undirected, where $[\![n]\!]$ denotes the set of integers $\{1, \dots, n\}$. We do not consider self-loops, as most applications do not contain self-interactions, but there is no hindrance to include them in the SBM if needed.

Denote $(\boldsymbol{\pi}, \boldsymbol{\gamma})$ the parameters of a SBM with K blocks, where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K) \in (0,1)^K$ are the block proportions with $\sum_{k \in [\![K]\!]} \pi_k = 1$ and $\boldsymbol{\gamma} = (\gamma_{k,l})_{k,l} \in (0,1)^{K \times K}$ is the connectivity matrix. If the network is undirected, $\boldsymbol{\gamma}$ must be symmetric. Let $\mathbf{Z} = (Z_1, \dots, Z_n) \in [\![K]\!]^n$ be a vector of independent discrete latent variables for the nodes, with $\mathbb{P}(Z_i = k) = \pi_k$ for all $k \in [\![K]\!]$ and $i \in [\![n]\!]$. When convient, we use the one-hot encoding $Z_i = (Z_{i,1}, \dots, Z_{i,K}) \in \{0,1\}^K$, where Z_i has multinomial distribution $\mathcal{M}(1, \boldsymbol{\pi})$. Conditionally on the node labels $\mathbf{Z}$, the observed adjacency matrix $A = (A_{i,j})_{1 \leq i,j \leq n} \in \{0,1\}^{n \times n}$ verifies

$$A|Z = \bigotimes_{(i,j) \in \mathcal{R}} A_{i,j}|Z_i, Z_j = \bigotimes_{(i,j) \in \mathcal{R}} \mathrm{Ber}\left(\gamma_{Z_i, Z_j}\right),$$

where $\mathrm{Ber}(\gamma)$ is the Bernoulli distribution. We then say that A has the distribution of a *(binary) stochastic block model* and denote $A \sim \mathrm{SBM}_n(\boldsymbol{\pi}, \boldsymbol{\gamma})$.

2.1 Dynamic stochastic block model

A recent branch of research is the statistical analysis of dynamic networks, as more and more data with recurrent interactions are available. We propose the first time-continuous extension of the popular stochastic block model for longitudinal networks by incorporating inhomogeneous Poisson processes that model the interaction events between pairs of nodes. The node clustering defined by the stochastic block model is used to reduce the number of unknown model parameters related to the Poisson processes. For the inference, a semiparametric variational Expectation-Maximization (EM) algorithm is developed, where the intensities of the Poisson processes are estimated in a nonparametric way, including adaptive model selection. This work is the fruit of a collaboration with Catherine Matias and Fanny Villers, published in [MRV18]. An implementation of the alogrithm is available via the R package `ppsbm` [GMRV18] and additional code is provided in [MRV18].

2.1.1 State of the art

The past years have seen a large increase in the interest for modelling dynamic interactions between individuals, as continuous-time information on interactions is now often available [3, 4]. However, most existing models are developed for a sequence of networks as the latent space joint model in [5]. In general, sequences of network snapshots are obtained by data aggregation over predefined time intervals (see [6] for a review). Discretization induces a loss of information, and so developing continuous-time models is an important issue.

The analysis of event data is an old area in statistics (see e.g. [7]). Generally, the number of interactions of all pairs (i, j) of individuals up to time t are modelled by a multivariate counting process $N(t) = (N_{i,j}(t))_{(i,j)}$. Various models use a set of statistics, that is chosen by the user, to modulate the interactions [8, 9, 10]. The choice of these statistics raises some issues: increasing their number may lead to a high-dimensional problem, and interpretation of the results might be blurred by their possible correlations. Other approaches aim at clustering time series or deriving a network that explains their coupling [11].

In statistical network analysis stochastic block models are widely used for several reasons. They easily accommodate any heterogeneous graph topology and parameters are interpretable which is important in applications. Furthermore, the SBM can be viewed as a dimension reduction technique, as nodes are clustered into groups of nodes with similar connecting behaviour (see the review [12]). For discrete-time sequences of graphs, recently several generalizations of the SBM to a dynamic context have been proposed [13, 14, 15, 16].

In our work we develop the first semiparametric SBM for continuous-time interaction events, where interactions are modelled by conditional inhomogeneous Poisson processes. In contrast to other approaches, our model does not use any predefined network statistics that modulate interactions, but intensities are modelled and estimated in a nonparametric way. Our estimation and clustering approach is a semiparametric version of the variational Expectation-Maximization (EM) algorithm based on nonparametric estimators of the intensities. Semiparametric generalizations of the classical EM algorithm have been proposed in other contexts, e.g. [17, 18, 19, 20].

However, we are not aware of other attempts to incorporate nonparametric estimates in a variational approximation algorithm.

2.1.2 Poisson-process stochastic block model

We observe the pairwise interactions of n individuals during the time interval $[0, T]$ given by

$$\mathcal{O} = \{(t_m, i_m, j_m), m \in [\![M]\!]\},$$

where (t_m, i_m, j_m) corresponds to the event that an interaction from individual $i_m \in [\![n]\!]$ to individual $j_m \in [\![n]\!]$ occurs at time $t_m \in [0, T]$.

As in the classical SBM, we assume that individuals belong to one out of K blocks according to block probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, encoded by the latent variables $\mathbf{Z} = (Z_1, \dots, Z_n) \in [\![K]\!]^n$. Moreover, the relation between two individuals i and j is driven by their node labels Z_i and Z_j. That is, conditionally on $\mathbf{Z}$, we suppose that the stochastic process $N_{i,j}(\cdot)$ of interactions from i to j is an inhomogeneous Poisson process with intensity, say $\gamma^{(k,l)}(\cdot)$, given that $Z_i = k$ and $Z_j = l$. In other words, in the *Poisson-process stochastic block model* (PPSBM) the set of observations $\mathcal{O}$ is a realization of the multivariate counting process $\{N_{i,j}(\cdot)\}_{(i,j)\in\mathcal{R}}$ with conditional intensity process $\{\gamma^{(Z_i,Z_j)}(\cdot)\}_{(i,j)\in\mathcal{R}}$. The process $N_{i,j}$ is not a Poisson process, but a counting process with intensity $\sum_{(k,l)\in[\![K]\!]^2} \pi_k \pi_l \gamma^{(k,l)}$. We denote $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$ the infinite-dimensional parameter of the PPSBM.

We also propose a zero-inflated version of the PPSBM, by introducing null intensities with positive probability. This model accommodates sparse networks as often encountered in applications. The adaptation of the inference algorithm to the sparse setting is straightforward.

Under very reasonable assumptions, the PPSBM and the sparse PPSBM are shown to be identifiable.

2.1.3 Semi-parametric variational EM-algorithm

As a latent variable model inference in the SBM is involved. Since the work of Daudin *et al.* [21] EM algorithms using a variational approximation [22] in the E-step have been frequently used to approximate the maximum likelihood estimator. We follow this approach for the PPSBM.

Variational E-step

The E-step of the EM-algorithm requires the knowledge of the posterior distribution of the latent variables $\mathbb{P}_\theta(\mathbf{Z} \mid \mathcal{O})$, which is not tractable because the Z_i are not conditionally independent. Thus, we perform a variational approximation of $\mathbb{P}_\theta(\mathbf{Z} \mid \mathcal{O})$ as in [21] by a simpler distribution, namely by a factorized distribution $\mathbb{P}_{\boldsymbol{\tau}}$ of the form

$$\mathbb{P}_{\boldsymbol{\tau}}(\mathbf{Z} = (k_1, \dots, k_n) \mid \mathcal{O}) = \prod_{i\in[\![n]\!]} \mathbb{P}_{\boldsymbol{\tau}}(Z_i = k_i \mid \mathcal{O}) = \prod_{i\in[\![n]\!]} \tau^{i,k_i}, \quad (k_1, \dots, k_n) \in [\![K]\!]^n,$$

for parameters $\tau^{i,k}$. More precisely, we search the distribution $\mathbb{P}_{\boldsymbol{\tau}}$ minimizing the Kullback-Leibler divergence, that is, $\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} \mathrm{KL}(\mathbb{P}_{\boldsymbol{\tau}}(\cdot \mid \mathcal{O}), \mathbb{P}_\theta(\cdot \mid \mathcal{O}))$. In the PPSBM this minimization amounts to solve a fixed point equation, which can be done numerically in very short time. The variational parameters $\tau^{i,k}$ are estimates of the posterior probabilities of the node labels $\mathbb{P}_\theta(Z_i = k|\mathcal{O})$, and as such, they define a soft clustering of the nodes.

Nonparametric M-step

Roughly, in the M-step the problem consists in estimating the intensity $\gamma_{k,l}$ of the weighted cumulative process $N_{\mathbf{Z}}^{(k,l)} = \sum_{(i,j)\in\mathcal{R}} Z_{i,k} Z_{j,l} N_{i,j}$ for every $(k,l) \in [\![K]\!]^2$. However, this process is unobserved, as it depends on the latent variables $\mathbf{Z}$. The idea is to use the current variational parameters $\tau^{i,k}$ to construct an empirical counterpart of $N_{\mathbf{Z}}^{(k,l)}$ defined by

$$N^{(k,l)}(E) = \int_E \mathrm{d}N^{(k,l)}(s) = \sum_{m\in[\![M]\!]} \tau^{i_m,k}\tau^{j_m,l}\mathbb{1}_E(t_m), \quad \text{for any interval } E, \tag{2.1}$$

and to estimate the intensity of this process $N^{(k,l)}$. Alternatively to $N^{(k,l)}$, one could define a simpler approximative process based on a hard clustering of the nodes. That is, replacing the variational parameters $\tau^{i,k}$ in (2.1) with the MAP-estimators $\hat{Z}_{i,k} = \mathbb{1}\{\hat{Z}_i = \arg\max_{l\in[\![K]\!]}\{\tau^{i,l}\} = k\}$ reduces the number of terms in the sum, but also yields a coarser estimate than the one obtained with our approach. Similar to the difference of the EM- and the classification EM (CEM) algorithm, we shall achieve more accurate results by using $N^{(k,l)}$ as defined above, while the additional computational cost is not prohibitive here.

We develop two nonparametric approaches for the estimation of the intensity of $N^{(k,l)}$. The first is a nonparametric kernel method, which is suited to estimate smooth functions, but may suffer from boundary effects that deteriorate the estimation accuracy. The second approach is based on piecewise constant functions to estimate the intensity $\gamma^{(k,l)}$. Such a histogram approach has some advantages compared to the kernel method as we will see below. Concretely, we adapt the adaptive intensity estimator for the Aalen multiplicative intensity model proposed by [23] to the PPSBM. To choose an appropriate partition of the time interval $[0,T]$ for the histogram estimator, a penalized least-squares criterion is used, that measures the fit of a candidate histogram estimator and the process $N^{(k,l)}$. It is conceived in a classical way such that fine partitions are penalized and a bias-variance trade-off is achieved. For reasons of computational efficiency we focus on nested regular dyadic partitions denoted by $\mathcal{E}_d$ with 2^d intervals of length $T2^{-d}$ for $d > 1$. The final adaptive intensity estimator has the simple form

$$\hat{\gamma}_{\mathrm{hist}}^{(k,l)}(t) = \frac{2^{\hat{d}^{(k,l)}}}{TY^{(k,l)}} \sum_{E\in\mathcal{E}_{\hat{d}^{(k,l)}}} N^{(k,l)}(E)\mathbb{1}_E(t),$$

where $Y^{(k,l)} = \sum_{(i,j)\in\mathcal{R}} \tau^{i,k}\tau^{j,l}$ is the estimated number of dyads (i,j) with latent groups (k,l) and the parameter $\hat{d}^{(k,l)}$ designates the optimal partition for the estimation of $\gamma^{(k,l)}$. We highlight that at every M-step, that is, at every iteration of the EM-algorithm, and for every pair (k,l), we apply the device for the optimal choice of the partition, that is of $\hat{d}^{(k,l)}$.

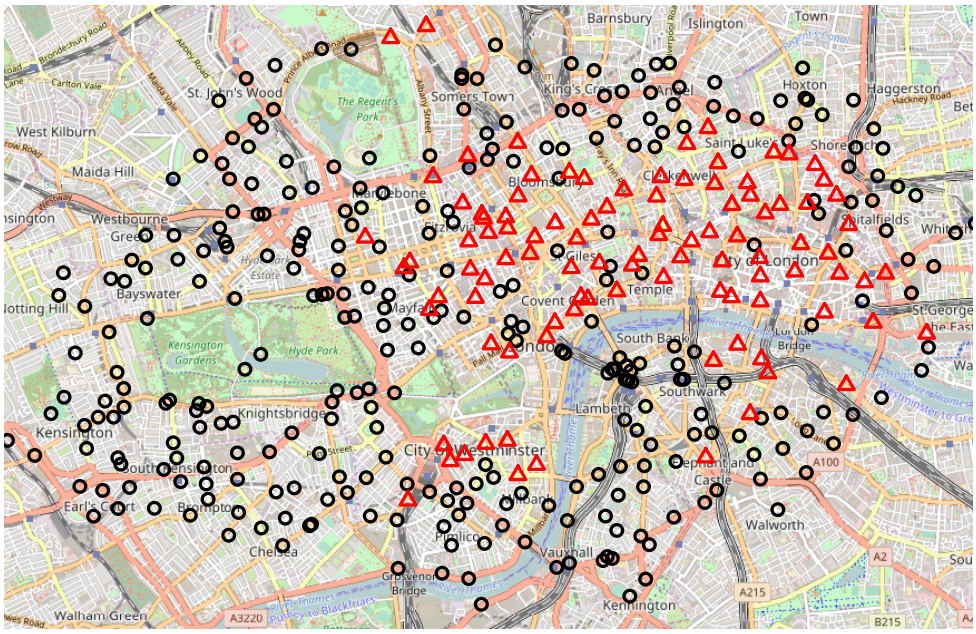


Figure 2.1: London bike stations and clustering into two clusters (represented by different colors) obtained with the sparse PPSBM. (Figure produced with OpenStreetMap project).

2.1.4 Model selection

In any SBM and its variants, the selection of the adequate number of blocks K is an issue. In [21], the integrated classification likelihood (ICL) criterion, which was first introduced in the mixture context in [24], has been adapted for model selection in the binary SBM. Roughly, in the context of model selection, the ICL is the complete-data variational log-likelihood penalized by the number of parameters. As in the PPSBM the parameter contains a nonparametric, that is, a infinite-dimensional part. So we need a trick to make the ICL work here. In fact, in the histogram approach, once the partitions $\mathcal{E}_{\hat{d}^{(k,l)}}$ are selected, there are only a finite number of parameters, namely $2^{\hat{d}^{(k,l)}}$ for every (k,l), to estimate. This yields the ICL criterion

$$\mathrm{ICL}(K) = \log \mathbb{P}_{\hat{\theta}(K)}(\mathcal{O}, \hat{\boldsymbol{\tau}}(K)) - \frac{1}{2}(K-1)\log n - \frac{1}{2}\log(|\mathcal{R}|) \sum_{(k,l)\in[\![K]\!]^2} 2^{\hat{d}^{(k,l)}},$$

where $\hat{\theta}(K)$ and $\hat{\boldsymbol{\tau}}(K)$ are the parameter estimates and variational parameters provided at the end of the algorithm when run with K blocks. The best number of blocks $\hat{K}$ is the one that maximizes the ICL, that is, $\hat{K} = \arg\max_{K\in[\![K_{\max}]\!]} \mathrm{ICL}(K)$ for some predefined upper bound $K_{\max}$. As kernel estimators cannot be parametrized by a finite-dimensional parameter, the ICL criterion is not defined and so this model selection device is available only in the histogram approach.

2.1.5 London bike sharing data

We illustrate the PPSBM on cycle hire usage data from the bike sharing system of the city of London [25]. Data consist in pairs of stations associated with a single hiring/journey (departure

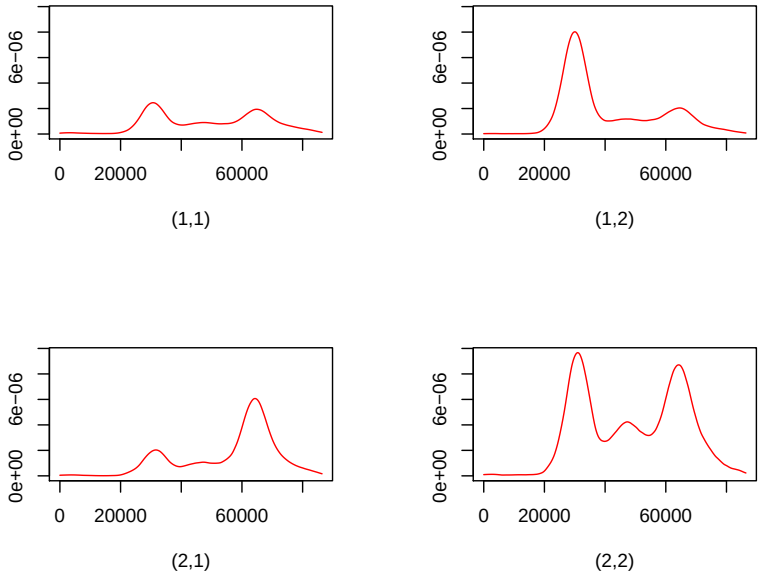


Figure 2.2: London bike sharing system: estimated intensities in the sparse PPSBM (time on the x-axis is in seconds).

station, ending station) and corresponding time stamp (hire time). For a randomly chosen day (1st February 2012) the dataset contains $n = 415$ stations and $M = 17,631$ hire events.

Our algorithm provides a PPSBM with $\hat{K} = 6$ latent blocks. An inspection of the clusters reveals that the blocks are indeed geographic clusters. This makes sense, as interacting stations are expected to be geographically close. Indeed, the dataset does not contain a single journey from one end of London to the other. Several of the estimated intensities are quite similar, so that we conclude that the clustering is mainly driven by the geographic locations of the bike stations.

Now, data are very sparse, as only a very small fraction (7%) of the processes $N_{i,j}$ are non null, that is contain at least one hiring event. So it makes sense to apply the sparse version of the PPSBM to the same data and interestingly only $\hat{K} = 2$ clusters are selected. The clustering is represented on a map in Figure 2.1. One group contains the central part of the city (red), while the remaining stations form a large peripheral cluster (black). According to the estimated intensities in Figure 2.2, the central group (group 2) has large intra-group intensity with three modes: one in the morning, at lunch and at the end of the day. The peripheral cluster (group 1) mostly consists in 'leaving' stations in the morning (with a mode in the morning in the intensity for $(k, l) = (1, 2)$) and in 'arriving' stations at the end of the day (with a mode in the intensity for $(k, l) = (2, 1)$). We conclude that, on the contrary to the first PPSBM, this clustering is driven by the different interaction behaviours and not by geographic locations. Both models, PPSBM and its sparse version, provide interesting results and are complementary as they shed different lights on the data.

2.2 Mini-batch sampling for scalable EM algorithms

Most standard EM-type algorithms are limited in the sample size they can deal with. This is also the case of the Monte Carlo Markov Chain Stochastic Approximation Expectation-Maximization (MCMC-SAEM) algorithm, an inference algorithm for general latent variable models including the SBM. We show how to scale the algorithm to large datasets by using mini-batch sampling. This is joint work with Estelle Kuhn and Catherine Matias. We published a paper [KMR20] and provide the code [MKR20].

2.2.1 State of the art

To speed up computing in the classical Expectation-Maximization (EM) algorithm [26] and its variants, various mini-batch [27, 28, 29, 30] and online versions [31, 32, 33, 34] have been proposed. They all consist in using only a part of the data during one iteration in order to shorten computing time and accelerate convergence. While online algorithms process a single observation per iteration handled in the order of arrival, mini-batch algorithms use larger, randomly chosen subsets of observations. The size of these subsets of data is generally called the mini-batch size. Choosing large mini-batch sizes entails long computing times, while very small mini-batch sizes and online algorithms may result in a loss of accuracy of the algorithm. This raises the question about the optimal mini-batch size that would achieve a compromise between accuracy and computing time. However this issue is generally overlooked.

2.2.2 Latent variable models and algorithm

We consider a common latent variable model with incomplete (observed) data $\mathbf{Y} \in \mathbb{R}^m$ and latent (unobserved) variable $\mathbf{Z}$. Denote n the dimension of the latent variable $\mathbf{Z} = (Z_1 \dots, Z_n) \in \mathbb{R}^n$. In many models, n also corresponds to the number of observations m, but this is e.g. not the case in the SBM, which is covered by our framework. Let $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ be the model parameter. We assume a general exponential model, that is, the complete-data likelihood function has the form $f(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta}) = \exp\{-\psi(\boldsymbol{\theta}) + \langle S(\mathbf{Y}, \mathbf{Z}), \phi(\boldsymbol{\theta})\rangle\} c(\mathbf{Y}, \mathbf{Z})$, where $S(\mathbf{Y}, \mathbf{Z}) \in \mathcal{S}$ is a vector of sufficient statistics. As the data $\mathbf{Y}$ are considered to be fixed realizations, we lighten notations by omitting all dependencies in $\mathbf{Y}$. That is, we write, for instance, $S(\mathbf{Z})$ instead of $S(\mathbf{Y}, \mathbf{Z})$.

Mini-batch MCMC-SAEM algorithm

The E-step of the traditional EM-algorithm consists in computing the conditional expectation $\mathbb{E}_{\boldsymbol{\theta}_{k-1}}[S(\mathbf{Z})]$ of the sufficient statistic under the current parameter value $\boldsymbol{\theta}_{k-1}$. When this expectation has no closed-form expression, it can be estimated by a stochastic approximation algorithm as done in the original MCMC-SAEM algorithm [35]. This means that the E-step is replaced with a simulation step using a MCMC procedure, namely a Metropolis-Hastings-within-Gibbs algorithm [36], combined with a stochastic approximation step.

When the dimension n of the latent variable $\mathbf{Z}$ is large, the simulation step can be very time-consuming as *all* latent components Z_i are simulated at *every* iteration. Thus, according

Algorithm 1 Mini-batch MCMC-SAEM

Input: Data $\mathbf{Y}$, mini-batch proportion α, step sizes $(\gamma_k)_{k\geq 1}$.
Initialization: Choose initial values $\boldsymbol{\theta}_0$, $\mathbf{S}_0$, $\mathbf{Z}_0$ for the model parameter, the sufficient statistic and the latent variable.
Set $k = 1$.
while not converged **do**
  Sample the number of latent components to be updated: $r_k \sim \text{Bin}(n, \alpha)$.
  Sample r_k indices from $[\![n]\!]$, denoted by $\mathcal{I}_k$.
  Set $\mathbf{Z}_k = \mathbf{Z}_{k-1}$
  for $i \in \mathcal{I}_k$ **do**
    Sample from the Metropolis kernel that acts only on the i-th coordinate:
      $\mathbf{Z} \sim \Pi_i(\mathbf{Z}_k, \cdot|\boldsymbol{\theta}_{k-1})$.
    Update latent variables: $\mathbf{Z}_k = \mathbf{Z}$.
  end for
  Evaluate the sufficient statistics $S(\mathbf{Z}_k)$ by a clever update of its previous value $S(\mathbf{Z}_{k-1})$.
  Perform the stochastic approximation step: $\mathbf{S}_k = (1-\gamma_k)\mathbf{S}_{k-1} + \gamma_k S(\mathbf{Z}_k)$.
  Update the model parameter by a classical M-step: $\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}(\mathbf{S}_k)$.
  Increment k.
end while

to the spirit of other mini-batch algorithms, updating only a part of the latent components may speed up the computing time and also the convergence of the algorithm. Denote $\alpha \in (0, 1)$ the average proportion of components of the latent variable $\mathbf{Z}$ that are updated during one iteration. Then in the E-step, we first randomly choose $\alpha 100\%$ of the latent components Z_i, which are then updated by sampling from the associated Metropolis kernel.

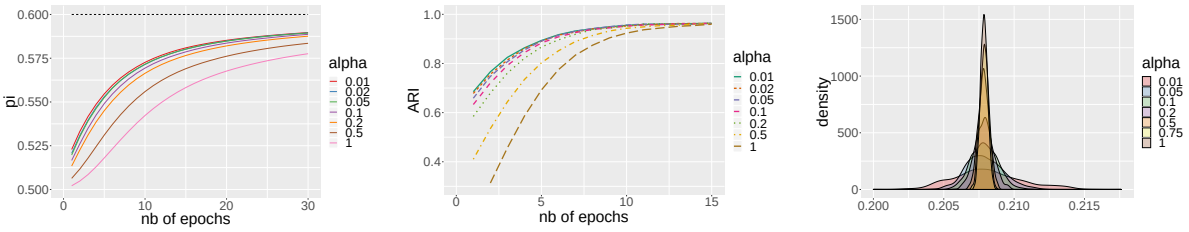
The naive evaluation of the sufficient statistic $S(\mathbf{Z}_k)$ on large datasets is too time-consuming. Though, in most models it is computationally much more efficient to derive the value of $S(\mathbf{Z}_k)$ from its previous value $S(\mathbf{Z}_{k-1})$ by correcting only for the terms that involve recently updated latent components. In general, this amounts to using only a small part of the data $\mathbf{Y}$ and thus speeds up computing. Algorithm 1 gives a complete description of the algorithm.

2.2.3 Convergence result

In the classical MCMC-SAEM algorithm (also called batch algorithm), the transition kernel describing the simulation step is a composition of n kernels of the form $\Pi = \Pi_n \circ \cdots \circ \Pi_1$, where Π_i only acts on the i-th coordinate. Now, for the mini-batch algorithm we introduce the kernel $\Pi_{\alpha,i}$ defined as a mixture of the original kernel Π_i and the identity kernel Id given by

$$\Pi_{\alpha,i}(\mathbf{Z}, \mathbf{Z}'|\boldsymbol{\theta}) = \alpha\Pi_i(\mathbf{Z}, (Z_1, \ldots, Z_i', \ldots, Z_n)|\boldsymbol{\theta}) + (1-\alpha)\text{Id}(\mathbf{Z}, \mathbf{Z}').$$

Then, the mini-batch simulation step corresponds to generating a latent vector $\mathbf{Z}$ according to the Markov kernel $\Pi_\alpha = \Pi_{\alpha,n} \circ \cdots \circ \Pi_{\alpha,1}$. With this kernel at hand, it can be seen that the mini-batch MCMC-SAEM algorithm formally belongs to the family of MCMC-SAEM algorithms with a particular choice of the transition kernel. Moreover, under assumptions that basically ensure convergence of the batch MCMC-SAEM algorithm, one can show the following convergence



(a) Parameter estimate $\hat{\pi}_1$. (b) ARI. (c) Histogram of final $\hat{\gamma}_{2,2}$.

Figure 2.3: Evolution of the parameter estimates of $\pi_1 = 0.6$ (a) and the ARI (b) with respect to the number of epochs. Limit distribution of the estimate of $\gamma_{2,2} = 0.2$ after 10 000 iterations for the mini-batch algorithm for mini-batch proportions $\alpha \in \{0.01, 0.05, 0.1, 0.3, 0.5, 0.8, 1\}$.

result.

Theorem 1. *Let $0 < \alpha \leq 1$ and $(\boldsymbol{\theta}_k)_{k\geq 1}$ be a sequence generated by the mini-batch MCMC-SAEM algorithm. Under appropriate model assumptions, almost surely,*

$$\lim_{k\to\infty} \boldsymbol{\theta}_k \in \{\boldsymbol{\theta} : \nabla \ell(\boldsymbol{\theta}) = 0\},$$

that is, $(\boldsymbol{\theta}_k)_{k\geq 1}$ converges to the set of critical points of the observed likelihood $\ell(\boldsymbol{\theta})$ as the number of iterations increases.

2.2.4 Speed up and accuracy

We conduct numerical experiments to assess the performance of the minibatch MCMC-SAEM algorithm in the stochastic block model. Figure 2.3 shows the typical evolution the parameter estimates (in (a)) and of the adjusted Rand index (ARI) (in (b)) along the algorithm for different mini-batch proportions. Estimates are compared in terms of epoch, where an epoch is the average number of iterations required to update n latent components. So Figure 2.3 (a) and (b) compare estimates for different mini-batch proportions at comparable computing time. We can see that the smaller the mini-batch proportion α, the faster the convergence, namely at the beginning of the algorithm. The fastest convergence is obtained with the smallest mini-batch proportion, which is characteristic for mini-batch sampling in any EM-type algorithm.

Let us give an intuitive explanation of this phenomenon. In general, the initial value $\boldsymbol{\theta}_0$ of the algorithm is far away from the target. So, during the first iteration of the batch algorithm, many time-consuming computations are done using the very bad value $\boldsymbol{\theta}_0$. Only at the very end of the first iteration, the parameter estimate is updated to a little better value $\boldsymbol{\theta}_1$. During the same time, a mini-batch algorithm with small α performs some computations with the same bad value $\boldsymbol{\theta}_0$, but reaches the M-step after a short time for the first update of $\boldsymbol{\theta}_0$. The new value $\boldsymbol{\theta}_1$ may be only a slight correction of $\boldsymbol{\theta}_0$, but, nevertheless, it is a move into the right direction and the next iteration is performed using a slightly better value than before. Metaphorically speaking, the batch algorithm makes long and time-consuming steps, but these steps are not necessarily directed into the best direction, whereas the mini-batch version makes plenty small

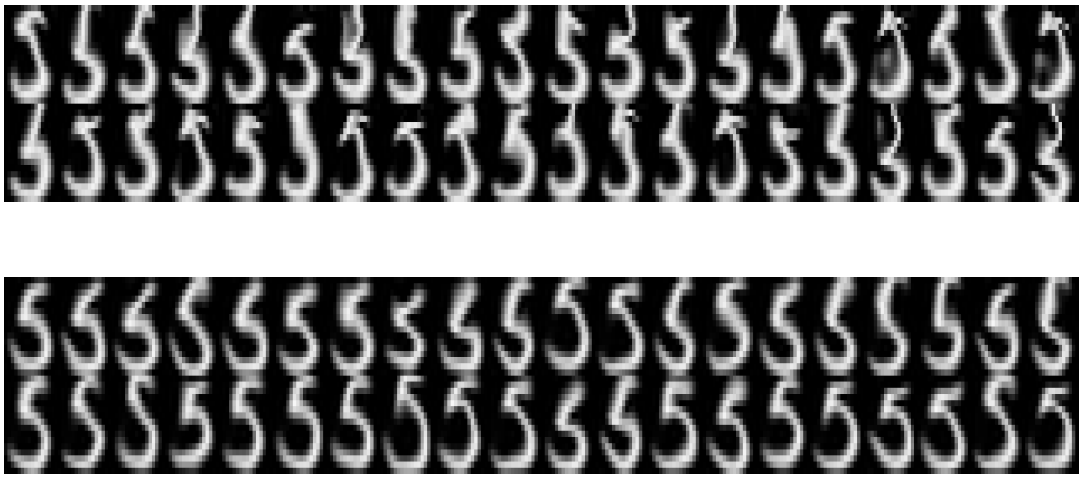


Figure 2.4: Synthetic images sampled from the model for digit 5 using the parameter estimates obtained with the batch version on 20 images (top) and with the mini-batch version with $\alpha = 0.2$ on 100 images (bottom).

and quick steps, correcting its direction after every step. As a whole, the mini-batch strategy results in a faster convergence of the algorithm as illustrated in Figure 2.3 (a) and (b).

Figure 2.3 (c) displays the histograms of the final estimates of parameter $\gamma_{2,2}$ obtained with the different mini-batch sizes for repeated runs of the algorithms. Contrary to the first simulations, here the comparison is done after 10 000 iterations (and not epochs), when all algorithms have attained convergence. We see that the histograms are approximately Gaussian, and interestingly, the variance increases when the mini-batch size α decreases. This is indeed coherent with the fact that in 10 000 iterations the batch version processes much more data than the mini-batch algorithms leading to more accurate estimation results. This raises the question on an optimal choice of the minibatch proportion α. It would be reasonable to seek a value of α that performs a trade-off between speeding up convergence and estimation accuracy.

2.2.5 Computing time constraints

To illustrate how to take advantage of the speed up of convergence in practice, consider the dense deformation template model [37], where observed images are assumed to be deformations of a common reference image. We use images of handwritten digits from the popular United States Postal Service database [38]. Suppose that we find the computing time acceptable when running the batch algorithm on $n = 20$ images during 1000 iterations, that is, until convergence. Can we do something better by using mini-batch sampling, say with $\alpha = 0.2$, within the same computing time? When applying the mini-batch algorithm on the same 20 images, convergence is attained much faster and there is a gain in computing time, but probably also some loss in accuracy. This is not what we are interested in, as we want to make use of the entire allotted computing time. Instead, we may increase the number of images in the input. As in the batch version 20 images are processed per iteration, the mini-batch algorithm with $\alpha = 0.2$ can be applied to a dataset with $n = 100$ images, since in average only 20 images are visited per iteration. Hence,

running the mini-batch algorithm with $\alpha = 0.2$ and $n = 100$ and the batch version with $n = 20$ over the same number of iterations takes roughly the same time. To assess the accuracy of the two procedures, we generate new samples from the learned models displayed in Figure 2.4. We see that the synthetic images generated with the mini-batch version (bottom row of the figure) resemble usual handwritten digits 5 more than the others. We conclude that, given a constraint on the computing time, more accuracy can be obtained by using the mini-batch MCMC-SAEM instead of the original algorithm.

2.3 Multiple testing related to graph inference

Jointly with my colleagues Etienne Roquain and Fanny Villers we use random graphs to address a multiple testing problem. For a set of entities, the goal is to test a null hypothesis for every pair of entities. By modelling the constellation of true and false null hypotheses by a stochastic block model and the observations as a noisy version of this graph, we recast the problem as a graph inference problem and successfully adapt the testing procedure of Sun and Cai [39]. The work is published in [RRV22], the accompanying R package is called `noisySBM` [RV20] and the code for the reproducibility of the results is available [RRV20].

2.3.1 State of the art

Paired null hypotheses occur in a large variety of domains such as social, biological or information sciences [40, 41, 42], typically in the form of scores describing interactions or similarity of pairs of entities. Thus, data have matrix form, potentially with large dimension. While the case of vector-based multiple testing inference is ubiquitous, matrix-based datasets are far less understood in the statistical literature. To the best of our knowledge, it has only been studied when the matrix is built upon pairwise comparisons between coordinates of the same observed vector, as it is the case with marginal or partial correlations, see [43, 44, 45] among others.

Our data are assumed to be directly collected in a matrix-wise fashion and the goal is to improve testing by incorporating structural information in the inference. For this, we follow the line of research based on the classical two-group mixture model introduced in [46]. The seminal works [47, 48] show how to control the false discovery rate while improving on Benjamini-Hochberg by consistently estimating the signal proportion, the null and alternative distributions. Further significant power enhancement can be obtained by incorporating some latent structure in the model, see [39] for group structure and [49, 50] for Markov structure.

We propose a random graph model that considers the observed matrice as a perturbation or a noisy version of an underlying binary graph. In existing models as in [51, 52, 53, 54, 55, 56] the uncertainty comes from a binary blurring mechanism of the underlying true network, that erroneously removes or adds edges according to some probabilities. On the contrary, we mainly work with Gaussian noise on the edges resulting in real-valued observations.

2.3.2 Noisy stochastic block model

Let n be the number of entities and $X = (X_{i,j})_{(i,j)\in\mathcal{A}}$ be a real-valued observed data matrix, where $X_{i,j} \in \mathbb{R}$ corresponds to a score, measurement or test statistic for the pair (i, j). We assume that a null hypothesis and an alternative is given for each pair and record the trueness/falseness of these null hypotheses in an unobserved binary matrix $A = (A_{i,j})_{(i,j)\in\mathcal{A}}$, for which $A_{i,j} = 0$ if and only if the null hypothesis for (i, j) is true. Specifically, this corresponds to the multiple testing setting where we test simultaneously for all $(i, j) \in \mathcal{A}$,

$$H_{0,i,j} : A_{i,j} = 0 \quad \text{versus} \quad H_{1,i,j} : A_{i,j} = 1.$$

The point of our work is to consider the binary matrix A as the adjacency matrix of a network, where edges correspond to false null hypotheses. Concretely, we model the distribution of A by a SBM, say $A \sim \mathrm{SBM}_n(\boldsymbol{\pi}, \boldsymbol{\gamma})$ with K blocks, block proportions $\boldsymbol{\pi}$, connectivity parameter $\boldsymbol{\gamma}$ and node labels $\mathbf{Z}$. The observation $X_{i,j}$ is obtained by replacing missing edges ($A_{i,j} = 0$) by pure random noise, that is, a realization of some null density g_{0,ν_0} with unknown parameter ν_0, whereas in place of present edges ($A_{i,j} = 1$) a signal is observed, modeled by an alternative density $g_{\nu_{k,\ell}}$ with parameters $\nu_{k,\ell}$ given that $Z_i = k$ and $Z_j = \ell$. The latter density depends on the block labels of the interacting nodes in the underlying SBM, such that the signal strength can be modulated locally. In short, conditionally on $\mathbf{Z}$ and A, in the *noisy stochastic block model* (NSBM) all $X_{i,j}$ are drawn independently as

$$X_{i,j}|\mathbf{Z}, A \sim (1 - A_{i,j})g_{0,\nu_0} + A_{i,j}g_{\nu_{Z_i,Z_j}}, \quad (i, j) \in \mathcal{A}.$$

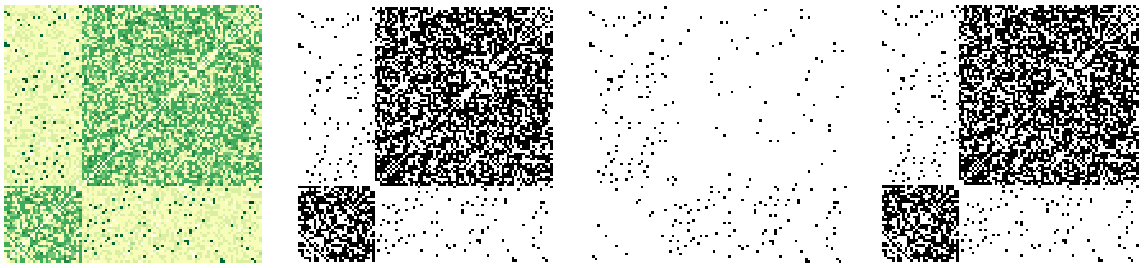
Our main example is the Gaussian case, where $\{g_{0,u}, u \in \mathcal{V}_0\} = \{\mathcal{N}(0, \sigma_0^2), \sigma_0 > 0\}$ and $\{g_u, u \in \mathcal{V}\} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$, which is particularly suitable when the observations $X_{i,j}$ correspond to test statistics that are known to be approximately Gaussian.

We state conditions under which the NSBM is identifiable. Furthermore, we propose a variational EM-algorithm for the inference. The algorithm provides an estimate $\hat{\theta}$ of all model parameters as well as estimates $\hat{\mathbf{Z}}$ of the node labels. The inference algorithm is implemented in the R packages `noisySBM` and it allows different types of distributions for the densities $g_{0,u}$ and g_u, namely Gaussian, exponential and Gamma distributions.

2.3.3 Multiple testing with structured ℓ-values

A multiple testing procedure is a measurable function $\varphi(X) \in \{0, 1\}^{\mathcal{R}}$ with $\varphi_{i,j}(X) = 1$ if and only if the null hypothesis on (i, j) is rejected. The associated *false discovery rate* (FDR) and *true discovery rate* (TDR) are defined by

$$\mathrm{FDR}(\varphi) = \mathbb{E}_\theta\left[\frac{\sum_{(i,j)\in\mathcal{R}}(1 - A_{i,j})\varphi_{i,j}(X)}{\sum_{(i,j)\in\mathcal{R}}\varphi_{i,j}(X)}\right], \quad \mathrm{TDR}(\varphi) = \frac{\mathbb{E}_\theta\left[\sum_{(i,j)\in\mathcal{R}} A_{i,j}\varphi_{i,j}(X)\right]}{\mathbb{E}_\theta\left[\sum_{(i,j)\in\mathcal{R}} A_{i,j}\right]}.$$



(a) Observation X. (b) Adjacency matrix A. (c) BH procedure. (d) New procedure.

Figure 2.5: (a) Real-valued observation matrix X with block structure. (b) Adjacency matrix A encoding the true (white) and false (black) null hypotheses. (c) Rejections of the Benjamini-Hochberg procedure. (d) Rejections of our new procedure based on a latent graph.

A multiple testing procedure is considered to be optimal when its FDR, that is, the average proportion of errors among the discoveries, is lower or equal some nominal level α, while its power, that is, the TDR, is as large as possible.

In the toy example in Figure 2.5, the observed matrix X in (a) comes from the Gaussian NSBM with underlying binary matrix A displayed in (b). Nodes are ordered according to the block labels of the nodes in the SBM revealing a strong underlying structure of the data. True null hypotheses give rise to pure noise in X (light-green points in (a)) and false null hypotheses result in signal (normal to dark-green points). As the signal strength depends on the node labels, here intermediate signal (normal green) is observed for intra-group edges, whereas the strongest signals occur for the few inter-group edges. The classical Benjamini-Hochberg (BH) procedure [57] is of the form $\varphi^{\mathrm{BH}}(X_{i,j}) = \mathbb{1}\{|X_{i,j}| > c\}$ for some threshold c. It is overly conservative as it only recovers a small part of the signal, see (c). The lack of power is due to the application of the same threshold c to all observations, failing to account for differences in the signal strength. A better strategy consists in choosing local thresholds that adapt to the data structure. In latent variable models, this can be achieved by considering the posterior probabilities, also called *structured ℓ-values*, given by $\ell_{i,j}(X, \mathbf{z}, \theta) = \mathbb{P}_\theta(A_{i,j} = 0 | X, \mathbf{Z} = \mathbf{z})$ instead of $|X_{i,j}|$. That is, the test procedure has the form

$$\varphi_{i,j} = \mathbb{1}\{\ell_{i,j}(X, \mathbf{Z}, \theta) \leq t\}, \tag{2.2}$$

for some threshold t, yielding much richer rejection regions than those of the form $\{|X_{i,j}| \geq c\}$ for $c > 0$. Rejection regions based on ℓ-values can be one-sided, two-sided with unbalanced sides and more. This results in a large gain of power, as one can see from Figure 2.5 (d), where the signal is almost perfectly recovered.

The threshold t can be chosen such that the so-called marginal FDR, given by $\mathrm{MFDR}_\theta(\varphi) = \mathbb{E}_\theta\left[\sum_{(i,j)\in\mathcal{R}}(1 - A_{i,j})\varphi_{i,j}(X)\right] / \mathbb{E}_\theta\left[\sum_{(i,j)\in\mathcal{R}}\varphi_{i,j}(X)\right]$, is controlled at level α. That is, threshold t is chosen such that $\mathrm{MFDR}_\theta(\varphi) = \alpha$. The MFDR is an approximation of the FDR and has the advantage to be handier and much easier to analyze than the FDR. The explicit computation of threshold t can then be circumvented by considering the the associated q-values.

2.3.4 Quasi-optimality of the procedure

We denote by φ^* the oracle procedure, which is the test procedure defined in (2.2) based on the true model parameter θ^*, the true node labels $\mathbf{Z}$ and the threshold t chosen as described above. A data-driven procedure denoted by $\widehat{\varphi}$ is obtained by plug-in, when the estimates $\hat{\theta}$ and $\hat{\mathbf{Z}}$ provided by the variational EM-algorithm are used in (2.2).

The oracle procedure φ^* is nearly optimal as stated in the following theorem. In particular, under mild model assumptions, the FDR is asymptotically controlled at level α. Moreover, among all procedures that control the MFDR, the oracle φ^* is the most powerful one.

Theorem 2 (Optimality of the oracle φ^*)**.** *Under appropriate assumptions,*

1. $\mathrm{MFDR}(\varphi^*) = \alpha$,
2. *for any procedure* φ *such that* $\mathrm{MFDR}(\varphi) \leq \alpha$, *it holds* $\mathrm{TDR}(\varphi^*) \geq \mathrm{TDR}(\varphi)$,
3. $\limsup_n \mathrm{FDR}\,(\varphi^*) \leq \alpha$.

Now our main result is that the data-driven procedure $\widehat{\varphi}$ mimics the behavior of the oracle φ^*, both in terms of FDR and TDR, up to remainder terms. It is obvious that the performance of $\widehat{\varphi}$ heavily depends on the quality of the estimator $\widehat{\theta}$ and the clustering $\widehat{Z}$. To state our results on $\widehat{\varphi}$, we introduce the following risk probability defined for any $\varepsilon > 0$ by

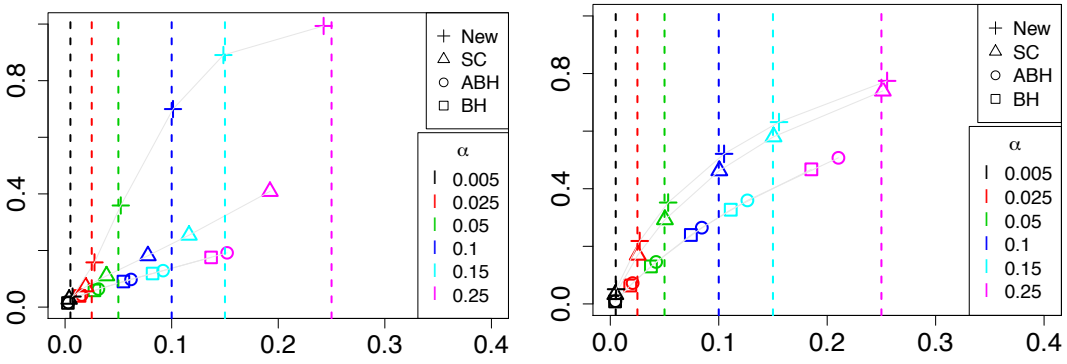
$$\eta(\theta^*, \varepsilon) = \mathbb{P}(\widehat{Z} \neq Z \text{ or } \|\widehat{\theta} - \theta^*\|_\infty > \varepsilon),$$

which corresponds to the probability that, either the clustering makes at least one mistake, or the estimator $\widehat{\theta}$ is more than ε away from the true parameter θ^*. Since the pioneer paper [58], several studies have suggested that in various SBM-type models, under appropriate restrictions on the parameter set Θ, the order of the risk probability $\eta(\theta^*, \varepsilon)$ becomes small when n increases [59, 60]. This is proved for the maximum likelihood estimator and alternatively for its variational approximation, and for a clustering based upon a maximum *a posteriori* approach, as used in our algorithm. We suppose that $\eta(\theta^*, \varepsilon)$ tends to 0 for any $\varepsilon > 0$ in the NSBM, but it is not proven explicitly.

Theorem 3 (Consistency of $\widehat{\varphi}$)**.** *Under appropriate regularity assumptions, if* $\eta(\theta^*, \varepsilon)$ *tends to 0 for any* $\varepsilon > 0$, *then*

$$\limsup_n \mathrm{FDR}\,(\widehat{\varphi}) \leq \alpha, \qquad \liminf_n \{\mathrm{TDR}\,(\widehat{\varphi}) - \mathrm{TDR}\,(\varphi^*)\} \geq 0.$$

Theorem 3 is in line with the state-of-the-art consistency results for the FDR and TDR in structured latent variable models, see [48, 49, 39, 61]. Now, the following theorem provides insights on the rate of convergence of the procedure $\widehat{\varphi}$. We underline that these theoretical results are non-asymptotic with respect to the number of tests, which is new to our knowledge compared to the existing multiple testing literature for mixture models.



(a) Two communities. (b) Preferential attachment.

Figure 2.6: Plot of $(\mathrm{FDR}_\alpha, \mathrm{TDR}_\alpha)$ for BH, ABH, SC and the new procedure $\widehat{\varphi}$. Dashed lines represent the nominal levels α.

Theorem 4 (Convergence rate of $\widehat{\varphi}$)**.** *Under appropriate assumptions, there exist constant C such that for any n large enough and any sequence $\varepsilon_n \geq \sqrt{\frac{\log n}{n}}$,*

$$\mathrm{FDR}\left(\widehat{\varphi}\right) \leq \alpha + C\varepsilon_n + \eta(\theta^*, \varepsilon_n), \qquad \mathrm{TDR}\left(\varphi^*\right) \leq \mathrm{TDR}\left(\widehat{\varphi}\right) + C\varepsilon_n + \eta(\theta^*, \varepsilon_n).$$

2.3.5 Numerical performance

Simulation results, as those presented in Figure 2.6, illustrate that the new data-driven testing procedure $\widehat{\varphi}$ controls the FDR at the nominal level. Furthermore, $\widehat{\varphi}$ largely outperforms state-of-the-art methods like BH, an adaptive BH procedure (ABH) based on [62, 63] and a Sun and Cai procedure (SC), which is based on thresholding cumulative means of ℓ-values, which are computed by estimating the alternative density by a mixture distribution [48]. Generally speaking, numerical experiments support the validity of our approach and also demonstrate its robustness with respect to model assumptions.

2.4 Model-based graph clustering

Today, entire collections of networks emerge in many fields of application [64, 65, 66, 67]. When analyzing multiple networks, most questions are related to graph comparison. The focus of the present work is on clustering of networks that do not share the same set of vertices and may vary in size. The goal is a method that partitions the networks according to their topology. Note that here the term graph clustering refers to the clustering of entire networks and not to the clustering of nodes of a single network as often considered in the literature. We propose a model-based clustering approach and an agglomerative algorithm for the inference. The method is described in [Reb22] and the algorithm is available via the R package `graphclust` [Reb23]).

2.4.1 State of the art

Networks have complex structure and thus graph comparison is not trivial. Graph similarity or graph distances can be defined in many ways [68, 69, 70, 71, 72, 73, 74]. Comparison can also be performed by hypothesis tests as in [75]. A widespread approach is based on graph embeddings. A graph embedding is a low-dimensional vector representation encoding structural information of the network. When networks share the same nodes, it makes sense to use node embeddings as for the task to cluster nodes of a single network as done in [76]. With a vector-valued graph embedding at hand, graph clustering is easily performed using off-the-shelf machine learning algorithms. However, such approaches do not account for the estimation uncertainty, since all graph embeddings are exactly treated in the same way regardless of the size of the network.

To overcome this problem a model-based clustering approach may be used that models the uncertainty associated with the observations (see [77] for a review). A statistical model also has the advantage to provide a natural framework for model selection, that is, the automated choice of the best number of clusters. In [78, 79] mixture models for the graph clustering are developed, the first for networks with node correspondence, the second for networks that do not share the same nodes. To the best of our knowledge, our model is the first that applies to collections of networks that do not share the same vertices and can be both directed or undirected. Furthermore, our model is much easier to interpret than models based on graphon estimates [80, 79] .

2.4.2 Mixture of stochastic block models

To perform model-based clustering, we follow the long-standing tradition of using a mixture model. To that end, a model for the mixture components has to be chosen. Here it is important to distinguish two cases, namely whether the node set is the same for all observed networks or whether there is no correspondence between the vertices of one network and another. For the first case, some approaches have been explored, namely by modeling mixture components by a stochastic block model [78] or a generalized linear models [81], or by measurement error models, where networks are considered to be perturbations of some ground-truth graph [82, 83]. In our work we focus on the second case, where networks have different sets of vertices and we propose a new mixture model. As we desire an interpretable model, we opt to use the stochastic block model for the mixture components.

Formally, let $\mathcal{A} = \{A^{(m)}), m \in [\![M]\!]\}$ be a collection of M networks, where $A^{(m)} = (A^{(m)}_{i,j})_{1 \leq i,j \leq n^{(m)}} \in \{0,1\}^{n^{(m)} \times n^{(m)}}$ denotes the adjacency matrix of the m-th network. Networks may have different numbers $n^{(m)}$ of vertices. We introduce independent discrete latent variables $\mathcal{U} = (U^{(1)}, \ldots, U^{(M)}) \in [\![C]\!]^M$ defining a partitioning of the M networks into $C \geq 1$ clusters. Denote $p_c = \mathbb{P}(U^{(m)} = c), c \in [\![C]\!]$ the cluster proportions and $\mathbf{p} = (p_1, \ldots, p_C) \in (0,1)^C$. Now, let $(\boldsymbol{\pi}^{(c)}, \boldsymbol{\gamma}^{(c)}), c \in [\![C]\!]$ be parameters of C different SBMs. The associated numbers of blocks, say K_c, are not constrained to be equal. We assume that all networks in cluster c are independent realizations of the SBM with parameter $(\boldsymbol{\pi}^{(c)}, \boldsymbol{\gamma}^{(c)})$. That is, in the *mixture model of SBMs*,

conditionally on $\mathcal{U}$,

$$\mathcal{A}|\mathcal{U} = \bigotimes_{m\in[\![M]\!]} A^{(m)}|U^{(m)} = \bigotimes_{m\in[\![M]\!]} \mathrm{SBM}_{n^{(m)}}\left(\boldsymbol{\pi}^{(U^{(m)})}, \boldsymbol{\gamma}^{(U^{(m)})}\right).$$

Denote $\theta = \left(\mathbf{p}, \{(\boldsymbol{\pi}^{(c)}, \boldsymbol{\gamma}^{(c)}), c \in [\![C]\!]\}\right)$ the parameters of the mixture model, and note that θ is identifiable only up to label switching. That is, switching cluster labels always results in the same probability distribution of $\mathcal{A}$. In addition, in every SBM, the node labels are also identifiable only up to label switching.

2.4.3 Hierarchical agglomerative algorithm

In a mixture model the clustering task becomes an inference problem, since cluster labels correspond to the latent variables of the model. In general model-based clustering, EM-type algorithms [84], MCMC [85] and hierarchical algorithms [86] are traditionally used to jointly infer cluster labels and model parameters. In the case of graph clustering, for mixtures of networks with a constant node set, EM algorithms are developed [78, 81] as well as Gibbs samplers [82, 83]. They all have the disadvantage that the number of clusters must be set by the user.

For our model we develop a hierarchical agglomerative algorithm that starts from an oversegmented clustering with singleton clusters. That is, we initialize the algorithm by a mixture of M SBMs and set $U^{(m)} = m$ for $m \in [\![M]\!]$, that is, every network forms a cluster on its own. Then clusters are successively merged to larger clusters while optimizing some criterion. Following the line of research initiated by [87], we choose the *integrated classification likelihood* (ICL) as the objective defined as

$$\mathrm{ICL}(\mathcal{A},\mathcal{U},\mathcal{Z}) = \log(p(\mathcal{A},\mathcal{U},\mathcal{Z})) = \log\left(\int p(\mathcal{A},\mathcal{U},\mathcal{Z}|\theta)p(\theta)\mathrm{d}\theta\right),$$

where $p(\theta)$ is a prior on the model parameters. The values $(\hat{\mathcal{U}}, \hat{\mathcal{Z}})$ that maximize the ICL, that is, $(\hat{\mathcal{U}}, \hat{\mathcal{Z}}) = \arg\max_{\mathcal{U},\mathcal{Z}} \mathrm{ICL}(\mathcal{A},\mathcal{U},\mathcal{Z})$, are convenient estimates of the graph clustering and the node labels. At every iteration, we choose the pair of clusters that yield the largest increase of the ICL when merging them. Obviously, for an efficient implementation of the algorithm, it is crucial that the evaluation of the increase of the ICL of merging any two clusters is fast. We show that this is in fact the case when choosing an appropriate prior distribution $p(\theta)$ and we provide a number of details and hints for speeding up computation.

Interestingly, the algorithm provides a whole cluster hierarchy that can be visualized by a dendrogram and intermediate clusterings are easily inspected. As the criterion includes a penalization of the number of clusters, the algorithm automatically stops when any further cluster aggregation results in a deterioration of the objective. Thus, model selection is performed automatically.

Furthermore, merging two clusters raises an issue related to the non-identifiability of the block labels in the SBM. In fact, node labels in the two clusters may not refer to the same type of blocks, but in our algorithm, for a given cluster, node labels must designate the same

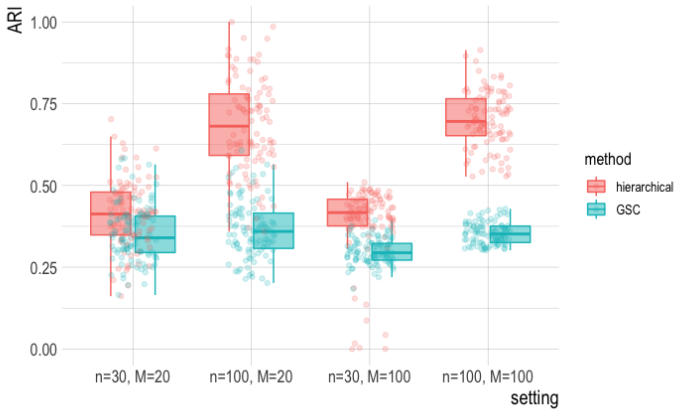


Figure 2.7: ARI for both our hierarchical algorithm and GCS for varying number of networks M and varying network sizes $n^{(m)}$. ARI computed over 100 datasets in each of the four settings.

SBM block in every network. If this is not the case, it is necessary to relabel the nodes when merging the clusters. A naive strategy consists in ordering one part of the SBM parameters, for instance, the block proportions $\pi_1, \dots, \pi_K$ or the diagonal elements of the connectivity matrix $\boldsymbol{\gamma}$ in a monotone order. However, as none of the parts of the parameter contains all relevant information, there are always cases where such an approach fails. To take into account both parts of the parameter $(\boldsymbol{\pi}, \boldsymbol{\gamma})$, we propose to use the graphon of the SBM.

The graphon, introduced by [88], is a function $g : [0,1]^2 \to [0,1]$ that can be used as a generative model for exchangeable random graphs including the SBM. First, generate independent random variables $U_i \sim U[0,1]$ for the vertices $i \in [\![n]\!]$. Then, conditionally on U_i and U_j, draw an edge $A_{i,j} \sim \mathrm{Ber}(g(U_i, U_j))$. The graphon of the model $\mathrm{SBM}_n(\boldsymbol{\pi}, \boldsymbol{\gamma})$ is given by

$$g_{(\boldsymbol{\pi},\boldsymbol{\gamma})}(u,v) = \gamma_{k,l} \qquad \text{for every } (u,v) \in R_{k,l} = (q_{k-1}, q_k] \times (q_{l-1}, q_l]\,, \tag{2.3}$$

where $q_k = \sum_{s \in [\![k]\!]} \pi_s$, $k \in [\![K]\!]$, $q_0 = 0$. Indeed, when $U_i \in (q_{k-1}, q_k]$, then $Z_i = k$. The graphon $g_{(\boldsymbol{\pi},\boldsymbol{\gamma})}$ is a piecewise constant function depending on the entire SBM parameter. Clearly, it also depends on the order of the block labels. Changing the block labels implies the permutation of the piecewise constant parts of the graphon.

To compare SBMs with parameters $(\boldsymbol{\pi}^{(c)}, \boldsymbol{\gamma}^{(c)})$ and $(\boldsymbol{\pi}^{(c')}, \boldsymbol{\gamma}^{(c')})$, consider the L^2-distance of their graphons. By the piecewise constant character, the squared distance is a finite sum given by

$$\begin{aligned} \|g_{(\boldsymbol{\pi}^{(c)},\boldsymbol{\gamma}^{(c)})} - g_{(\boldsymbol{\pi}^{(c')},\boldsymbol{\gamma}^{(c')})}\|_2^2 &= \int_{[0,1]^2} (g_{(\boldsymbol{\pi}^{(c)},\boldsymbol{\gamma}^{(c)})}(u,v) - g_{(\boldsymbol{\pi}^{(c')},\boldsymbol{\gamma}^{(c')})}(u,v))^2 \mathrm{d}(u,v) \\ &= \sum_{k,l,k',l'} \left(\gamma^{(c)}_{k,l} - \gamma^{(c')}_{k',l'}\right)^2 |R_{k,l,k',l'}|, \end{aligned} \tag{2.4}$$

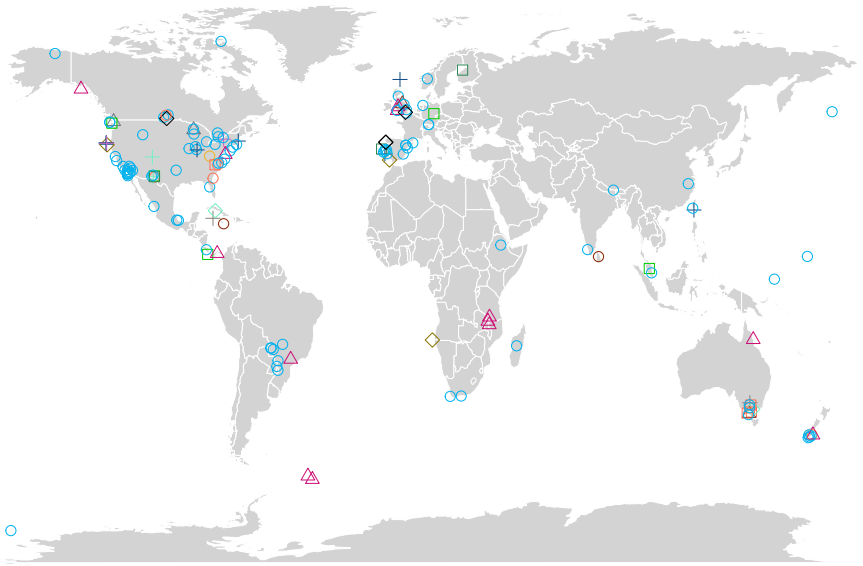


Figure 2.8: Geographical representation of the clustering of the foodwebs.

where $|R_{k,l,k',l'}|$ denotes the area of $R_{k,l,k',l'}$ defined as

$$R_{k,l,k',l'} = \left\{\left(\pi_{k-1}^{(c)}, \pi_k^{(c)}\right] \cap \left(\pi_{k'-1}^{(c')}, \pi_{k'}^{(c')}\right]\right\} \times \left\{\left(\pi_{l-1}^{(c)}, \pi_l^{(c)}\right] \cap \left(\pi_{l'-1}^{(c')}, \pi_{l'}^{(c')}\right]\right\}.$$

Roughly, our tool to match block labels of two SBM parameters consists in finding the permutations of the block labels yielding the smallest graphon distance. One can avoid the exploration of the set of all possible permutations by reordering the blocks of the graphon according to its marginals [89], rendering the evaluation of the graphon distance computationally fast. Note that the graphon distance in (2.4) is well-defined even when the number of blocks of the two models differ, that is, our tool can be used to compare two SBMs with different numbers of blocks. This tool may have an interest beyond the here considered clustering task and may be useful whenever SBMs have to be compared.

2.4.4 Properties of the clustering procedure

A numerical study highlights numerous properties of our algorithm. First, concerning estimation accuracy, it is well known that in the single network setting parameter estimates converge to the true SBM parameter when the number of nodes increases. Now, in the multiple network framework a different question is the accuracy of the estimators as a function of the number M of networks, when the network size $n^{(m)}$ is bounded. Fitting a standard SBM to a single small network yields SBM estimates with less blocks than in the true underlying model, since data do not provide enough evidence to estimate the parameters of a more complex model. However, we show that our approach that jointly analyzes all networks, allows to discover the true number of blocks and provides a better parameter estimate, even when all networks in the collection are small.

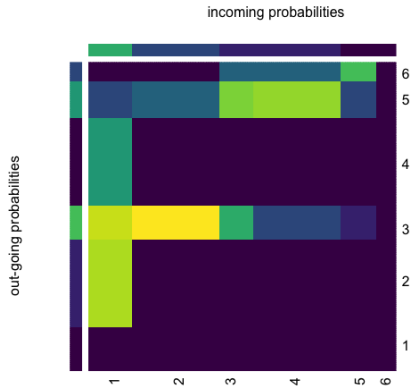


Figure 2.9: Graphon of SBM parameter of the dominant cluster with in-coming and out-coming probabilities on the sidebars.

Second, Figure 2.7 shows that increasing the number of vertices $n^{(m)}$ per network has not the same effect on the quality of the clustering as increasing the number of networks M. Collections that contain large networks provide better clustering results than those composed of small networks only.

Third, we compare our clustering algorithm to alternative methods. Most graph clustering procedures in the literature are based on a graph distance and apply spectral clustering to derive a clustering. Here we consider the (minimal) graphon distance defined in (2.3) to compare networks pairwisely and refer to this approach as the graphon spectral clustering (GSC) method. Figure 2.7 illustrates that in the four settings GSC is largely outperformed by our graph clustering approach although the correct number of clusters was provided to the GCS method. Moreover, no substantial improvement of GSC is observed when more data are available. This is in accordance with our understanding of such approaches, where the estimation uncertainty is not taken into account and networks are only analyzed separately. We conclude that model-based approaches as ours, where a common descriptor of each cluster is computed using all data associated with the cluster have a real advantage over graph distance methods.

Fourth, our cluster algorithm is robust to model misspecifications. In particular, the algorithm is able to make a distinction between data from the mixture model and outliers.

2.4.5 Application to ecological networks

Finally, an application to ecological networks, the mangal data base provided by [67], highlights the interpretability of the model and its usefulness for ecology. Our algorithm clusters 187 foodwebs from all over the world into 17 clusters, see Figure 2.8, revealing a dominant structure shared by about two third of the species. The SBM parameter of this universal ogranization of a foodweb, represented in Figure 2.9, indicates that in this organization about 43% of the species are vegetarians, 18% are predators and the rest of the species is somewhere in the middle of the food pyramid with both good chances to be eaten and to eat others. The foodwebs in the other clusters have significantly different graph topology and a comparison of their SBM parameters

allows to appreciate the differences between the different foodweb structures.

Chapter 3

Nonparametric density estimation in inverse problems

Inverse problems are often motivated by applications, where the distribution of interest may not be observed directly but only through noise or with some nonlinear distortion. In this chapter, two specific inverse problems are considered: continuous mixture models, namely mixtures of exponential densities, and biased data models, where a known nonlinear transformation deforms the target distribution. We propose nonparametric density estimates that come with a sound theoretical foundation. Our methods rely on modern statistical techniques. In the first work we provide minimax bounds, in the second we derive oracle-type risk bounds for adaptive estimators and compare the numerical performance of different estimates for the same task.

3.1 Minimax estimation of a mixing density

In contrast to finite mixture models, in a continuous mixture every observation is generated with an individual parameter. That is, the distribution of the latent parameter is not discrete, but continuous, taking its values in an entire interval. We investigate the problem of estimating the mixing density from observations of such a continuous mixture. We have a special interest in scale mixtures like mixtures of exponential distributions, which are regularly encountered in physics, but rarely studied in the statistical literature. We also address the problem of estimating the support of the mixing density. This is joint work with François Roueff, that we started at the end of my PhD. It is published in [RR15].

3.1.1 State of the art

Continuous mixtures immerge naturally in many fields of application, whenever an individual parameter is associated with each observation. The mixture density π_f is defined as

$$\pi_f(x) = \int_\Theta f(t)\pi_\theta(x)\mathrm{d}\theta,$$

where $\{\pi_\theta, \theta \in \Theta\}$ is a collection of densities and f the so-called *mixing density*. When θ is a scale parameter the model is called a *scale mixture*. Scale mixtures of uniforms are related to multiplicative censoring introduced in [90] and length-biased data, see [91]. Exponential mixtures play a significant role in natural science phenomena of discharge or disexcitation. From a mathematical point of view, scale mixtures are particularly interesting as they define classes of densities that verify some monotonicity constraints: mixture of Beta distributions $B(1,k)$ are *k-monotone* densities, and any *completely monotone* function can be written as an exponential mixture [92].

To estimate the mixing density different estimation strategies have been explored. Kernel estimators were considered [93, 94, 95] in particular when θ is a location parameter [96]. For mixtures of discrete distributions, orthogonal series estimators have been developed and studied in [97] and [98] and shown to have similar or better rates of convergence than the kernel estimator in [99]. Other projection estimators are based on Laguerre functions [100] or the Mellin transform [101].

3.1.2 Orthogonal series estimator

Our goal is to identify the mixing density f when a sample $X_1, \dots, X_n$ of the continuous mixture π_f is observed. The mixing density f is assumed to be square integrable, that is, $f \in L^2[a,b]$. Let $(\psi_k)_{k\geq 1}$ be a complete orthonormal basis of $L^2[a,b]$, so that f can be written as the orthogonal series $f(\theta) = \sum_{k\geq 1} c_k \psi_k(\theta)$ with $c_k = \langle f, \psi_k \rangle$. So estimating the coefficients c_k yields an estimator of f. This can be achieved by using the following property that holds in any mixture model. The moments of the mixture distribution correspond to the inner product of the mixing density f with some function φ. More precisely, for any nonnegative integrable function g,

$$\mathbb{E}_{\pi_f}[g(X)] = \int g(x)\pi_f(x)\mathrm{d}x = \int f(\theta) \underbrace{\int g(x)\pi_\theta(x)\mathrm{d}x}_{=\varphi(\theta)} \mathrm{d}\theta = \langle f, \varphi \rangle.$$

It follows that, if we find functions $(g_k)_{k\geq 1}$ such that $(\varphi_k)_{k\geq 1}$ defined as $\varphi_k(\theta) = \mathbb{E}_{\pi_\theta}[g_k(X)]$ are linearly independent functions in $L^2[a,b]$, then a sequence of orthonormal functions $\psi_1, \psi_2, \dots$ can be easily constructed. That is, the functions ψ_k can be written as linear combinations $\sum_{j\in[\![k]\!]} Q_{k,j}\varphi_j$ with known coefficients $(Q_{k,j})_{j\in[\![k]\!]}$. This yields explicit moment estimates, say $\hat{c}_{n,k}$, of the coefficients c_k. Finally, choosing an approximation parameter m, an estimator of the mixing density f is given by $\hat{f}_{m,n} = \frac{1}{n}\sum_{k\in[\![m]\!]} \hat{c}_{n,k}\psi_k$. We refer to $\hat{f}_{m,n}$ as the *orthogonal series estimator* or the *projection estimator* of approximation order m, as it is an approximation of the projection of f on the subspace $V_m = \mathrm{span}(\varphi_1, \dots, \varphi_m)$.

Those functions g_k depend on the mixture model, that is, on the collection $\{\pi_\theta, \theta \in \Theta\}$. For exponential distributions, that is, when $\pi_\theta(x) = \theta \mathrm{e}^{-\theta x}$, one can choose $g_k(x) = \mathbb{1}\left\{x > k - \frac{1}{2}\right\}$ for $k \geq 1$, yielding $\varphi_k(\theta) = \mathrm{e}^{-(k-\frac{1}{2})\theta}$.

Now, in many examples, there exists an operator T that transforms every function φ_k into a polynomial of degree $k-1$. In the example above on exponential mixtures, such an operator is $Tf(t) = f(-\log t)/\sqrt{t}$ for $t \in [\mathrm{e}^{-b}, \mathrm{e}^{-a}]$. By this trick, our estimator can be considered as a

polynomial approximation of Tf in some auxiliary space $L^2[a', b']$. In the following study of the estimator we suppose that the coefficients $Q_{k,j}$ are those of normalized Legendre polynomials in the space $L^2[a', b']$.

3.1.3 Rates of convergence and minimax risk

We provide an analysis of the orthogonal series estimator. First, we analyze the mean integrated squared error, then in the exponential mixture case, the estimator is shown to achieve the optimal minimax rate.

Mean integrated squared error

The mean integrated squared error (MISE) of our estimator has the classical bias-variance decomposition.

Proposition 1. *The orthogonal series estimator $\hat{f}_{m,n}$ satisfies*

$$\mathbb{E}\left\|\hat{f}_{m,n} - f\right\|^2 = \|P_{V_m} f - f\|^2 + \frac{1}{n}\mathrm{tr}\left(Q\Sigma Q^T\right),$$

where P_{V_m} is the projection operator on $V_m = span(\varphi_1, \ldots, \varphi_m)$, Σ is the covariance matrix of $(g_1(X_1), \ldots, g_m(X_1))$ and Q is the matrix containing all coefficients $Q_{k,j}$.

From the squared bias $\|P_{V_m} f - f\|^2$ we see that the performance of the estimator depends on how well f can be approximated by functions in V_m. To be more precise, for any approximation rate index α and radius C, define the approximation class

$$\mathcal{C}(\alpha, C) = \{f \in L^2[a, b] : \|f\| \leq C \text{ and } \|P_{V_m} f - f\| \leq C m^{-\alpha} \text{ for all } m \geq 1\}.$$

So if the mixing density f belongs to $\mathcal{C}(\alpha, C)$, then the bias is well controlled, namely it decreases at the rate $m^{-\alpha}$ as m increases. Now using properties of the Legendre polynomials, one obtains the following convergence rates of the MISE.

Theorem 5. *Suppose that $f \in \mathcal{C}(\alpha, C)$. Let $\hat{f}_{m,n}$ be the orthogonal series estimator with Legendre polynomials coefficients. Then*

$$\mathbb{E}\left\|\hat{f}_{m_n,n} - f\right\|^2 \leq C^2 m_n^{-2\alpha}(1 + o(1)),$$

in either of the following two cases:

(i) with $m_n = A \log n$ with appropriate constant A and if there are constants B, C_0 such that $\mathrm{Var}(g_k(X)) < C_0 B^{2k}$.

(ii) with $m_n = A \log n / \log\log n$ with appropriate constant A and if there are constants C_0, η such that $\mathrm{Var}(g_k(X)) < C_0 k^{\eta k}$.

Consequently, according to the considered case, the estimator $\hat{f}_{m_n,n}$ achieves the MISE rates $(\log n)^{-2\alpha}$ and $(\log(n)/\log\log n)^{-2\alpha}$ uniformly on the set of densities in $\mathcal{C}(\alpha, C)$. Concerning

the estimator of our example in the exponential mixture, one can show that condition (i) is satisfied and thus the convergence rate is of order $(\log n)^{-2\alpha}$. Below we will show that this is the optimal rate, that is, our estimator is minimax.

Approximation classes

Although the approximation classes $\mathcal{C}(\alpha, C)$ appear naturally in our study, they are not very intuitive and depend on the chosen functions g_k. Therefore, we show that the classes $\mathcal{C}(\alpha, C)$ are equivalent to more common smoothness classes. To that end, consider the weighted moduli of smoothness, denoted by $\omega_\varphi^r(f,t)_p$, that were introduced by [102] for the study of the rate of polynomial approximations. For constants $\alpha > 0$ and $C > 0$, we define the following class of functions in $L^2[a,b]$

$$\tilde{\mathcal{C}}(\alpha, C) = \{f \in L^2[a,b] : \|f\| \leq C \text{ and } \omega_\varphi^r(f,t)_2 \leq Ct^\alpha \text{ for all } t > 0\},$$

where $\varphi(x) = \sqrt{(x-a)(b-x)}$ and $r = [\alpha] + 1$. The following theorem states the equivalence of the classes $\mathcal{C}(\alpha, C)$ and $\tilde{\mathcal{C}}(\alpha, C)$.

Theorem 6. *Let $\alpha > 0$. Suppose that the opereator T has the form $Tg = \sigma \times g \circ \tau$ with sufficiently smooth functions σ and τ. Then there are constants C_1 and C_2 such that for all $C > 0$*

$$\mathcal{C}(\alpha, C_1 C) \subset \tilde{\mathcal{C}}(\alpha, C) \subset \mathcal{C}(\alpha, C_2 C).$$

This means that the classes $\mathcal{C}(\alpha, C)$ are equivalent to classes defined using weighted moduli of smoothness. This, in turn, relates them to Sobolev and Hölder classes.

Minimax rates

We study the question of the best possible convergence rate. In other words, we search a lower bound of the minimax risk defined as

$$\inf_{\hat{f} \in \mathcal{S}_n} \sup_{f \in \mathcal{C}} \mathbb{E}_{\pi_f} \|\hat{f} - f\|^2,$$

where $\mathcal{S}_n$ is the set of all Borel functions from $\mathbb{R}^n$ to $L^2[a,b]$, and $\mathcal{C}$ denotes a subset of densities in $L^2[a,b]$. In [RR15] we first provide a rather general lower bound of the minimax risk that covers a large spectrum of mixture models. However, to exhibit convergence rates for a specific mixture, (much) more work is required. In our paper the specific cases of exponential mixtures, Gamma shape mixtures and mixtures of compactly supported scale families are investigated in more detail. Our main result concerns exponential mixtures and we prove the following minimax lower bound.

Theorem 7. *In the exponential mixture model, there exists a constant C^* such that*

$$\inf_{\hat{f} \in \mathcal{S}_n} \sup_{f \in \tilde{\mathcal{C}}(\alpha, C)} \mathbb{E}_{\pi_f} \|\hat{f} - f\|^2 \geq C^* (\log n)^{-2\alpha} (1 + o(1)).$$

This yields that the orthogonal series estimator with the functions g_k as chosen above in the example and $m_n = O(\log n)$ is optimal in the sens that it achieves the minimax rate. Moreover, in the case of Gamma shape mixtures we show that our orthogonal series estimator achieves the minimax rate up to a $\log\log n$ multiplicative term.

3.1.4 Support estimation

A basic assumption of our estimation approach is that the mixing density f belongs to $L^2[a,b]$. However, in practice the exact interval $[a,b]$ is generally unknown. To compass this problem, we propose an estimator of the support of the mixing density f. Again we take advantage of polynomial approximations. Note that the problem of having to deal with an unknown support also occurs in classical density estimation, see [103, 104, 105].

Our support estimator is more precisely an estimator of the support of Tf denoted by $[a_0, b_0]$ and the idea is to consider as estimates the smallest and largest value where the estimator $T\hat{f}_{n,m_n}$ exceeds some threshold $\varepsilon_n/2$, by disregarding side-effects of size η_n. More precisely,

$$\hat{a}_n = \inf\left\{u \in [a',b'] : T\hat{f}_{n,m_n}(v) > \frac{\varepsilon_n}{2} \text{ for all } v \in [u, u+\eta_n]\right\}$$
$$\hat{b}_n = \sup\left\{u \in [a',b'] : T\hat{f}_{n,m_n}(v) > \frac{\varepsilon_n}{2} \text{ for all } v \in [u-\eta_n, u]\right\}.$$

If Tf decrease fast enough to zero on the borders of the interval $[a_0, b_0]$, then we can show that our support estimator is consistent for a convenient choice of the sequences $(\varepsilon_n)_n$ and $(\eta_n)_n$.

Proposition 2. *For sequences $m_n \to \infty$, $\varepsilon_n \to 0$ and $\eta_n \to 0$ such that $\mathbb{E}\left\|\hat{f}_{n,m_n} - f\right\|_{\mathbb{H}}^2 = O\left(m_n^{-2\alpha}\right)$, $\varepsilon_n^{-1} = o\left(m_n^{(2\alpha-1)/(2+1/\alpha')}\right)$ and $\eta_n = O\left(\varepsilon_n^{1/\alpha'} m_n^{-1}\right)$, the estimators $\hat{a}_n$ and $\hat{b}_n$ are consistent for the support bounds a_0 and b_0. More precisely, as $n \to \infty$,*

$$(\hat{a}_n - a_0)_+ = O_P\left(\varepsilon_n^{1/\alpha'}\right) \quad \text{and} \quad (\hat{a}_n - a_0)_- = O_P\left(\varepsilon_n^{1/\alpha'} m_n^{-1}\right),$$
$$(\hat{b}_n - b_0)_+ = O_P\left(\varepsilon_n^{1/\alpha'} m_n^{-1}\right) \quad \text{and} \quad (\hat{b}_n - b_0)_- = O_P\left(\varepsilon_n^{1/\alpha'}\right).$$

3.2 Adaptive estimation in biased data models

In various applications, observations are not directly available from the target distribution due to noise, missing data, censored or truncated observations. Those nonlinear distortions yield specific inverse problems and make functional estimation difficult. My interest for such models goes back to my doctoral thesis that was concerned with the so-called pile-up model and for which I proposed new methods in a parametric setting [RRS10, RRS11].

With Fabienne Comte we have then taken up the challenge of developing new nonparametric density estimators that best address the constraints encountered in practice. Based on modern techniques of nonparametric estimation we explore different strategies to correct nonlinear deformations in different contexts. We also provide oracle-type risk bounds for the mean integrated squared error (MISE) of the proposed adaptive estimators. Extensive numerical experiments complete the study. The work is published in [CR12] and [CR16].

3.2.1 The pile-up model and a general biased data model

Two models are studied, where the first, the so-called pile-up model, is a special case of the more general biased data model.

Pile-up model

The pile-up model is encountered in time-resolved fluorescence, where the fluorescence lifetime is the duration that a molecule stays in an excited state before emitting a photon [106, 107]. The distribution of the fluorescence lifetimes associated with a sample of molecules provides precious information on the underlying molecular processes. They are used by chemists to determine, for instance, the speed of rotating molecules or molecular distances.

Measurements are obtained by a technique called Time-Correlated Single-Photon Counting (TCSPC) [108]. After exciting a random number of molecules by a laser pulse, only the arrival time of the fastest photon striking the detector is observed. In other words, in the *pile-up model* an observation is defined as the minimum of a random number of independent and identically distributed (iid) variables following the target distribution. That is, observations $Z_1, \dots, Z_n$ are given by

$$Z_i = \min\{X_{i,1}, \dots, X_{i,N_i}\}, \quad i \in [\![n]\!], \tag{3.1}$$

where $(X_{i,k})_{i,k\geq 1}$ are iid random variables with density f and cumulative distribution function (cdf) F, and the random variables $(N_i)_{i\geq 1}$ are iid with Poisson distribution $\mathcal{P}(\theta)$ restricted on $\{1, 2, \dots\}$ and independent from $(X_{i,k})_{i,k\geq 1}$. The aim is to recover the density f of the variables $X_{i,k}$ from the observations $Z_1, \dots, Z_n$ without knowledge of the numbers $(N_i)_{i\geq 1}$ over which the minimum is taken.

In TCSPC, the Poisson parameter θ is a tuning parameter chosen by the user. Recent studies have made it clear that from a statistical information viewpoint, it is preferable to operate TCSPC in a mode with considerable pile-up effect [RRS11]. Consequently, estimation procedures are required that take the pile-up effect into account.

In one of our works we consider the specific case, where $(X_{i,k})_{i,k\geq 1}$ are supposed to be independent copies of X defined as

$$X = Y + \eta, \quad \text{with} \quad Y \sim f, \;\; \eta \sim f_\eta, \;\; Y \perp\!\!\!\perp \eta. \tag{3.2}$$

Here, η represents some additional measurement error or noise attributed to the measuring instrument. Then we want to recover the density f of Y from observations $Z_1, \dots, Z_n$ given by (3.1).

Biased data model

In the pile-up model, the observed distribution G is the result of a nonlinear distortion of the target distribution F. More precisely, $G(z) = 1 - M_\theta \circ (1 - F)(z)$ with $M_\theta(u) = \mathbb{E}(u^N) = (\mathrm{e}^{\theta u} - 1)/(\mathrm{e}^\theta - 1)$. This can be generalized by introducing a general known link function H :

$[0,1] \to [0,1]$ and assuming that

$$G(z) = H \circ F(z), \qquad z \in \mathbb{R}. \tag{3.3}$$

We refer to this model as the *biased data model*. It holds that

$$f(z) = w \circ G(z)\ g(z), \quad z \in \mathbb{R} \quad \text{with} \quad w(u) = \frac{1}{H' \circ H^{-1}(u)}, \qquad u \in [0,1]. \tag{3.4}$$

It follows the fundamental property that, for any measurable bounded function ψ, we have

$$\mathbb{E}_{X\sim F}[\psi(X)] = \mathbb{E}_{Z\sim G}\left[\psi(Z)\ w \circ G(Z)\right]. \tag{3.5}$$

This relation is the basis for the construction of moment estimators of $\mathbb{E}[\psi(X)]$ based on a sample $Z_1, \dots, Z_n$ from the distorted distribution G. Replacing the cdf G by its empirical version $\hat{G}_n(z) = \frac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{1}\{Z_i \le z\}$ yields a natural estimator of $\mathbb{E}[\psi(X)]$ by

$$\hat{L} = \frac{1}{n}\sum_{i\in[\![n]\!]} \psi(Z_i)\ w \circ \hat{G}_n(Z_i) = \frac{1}{n}\sum_{i\in[\![n]\!]} \psi(Z_{(i)})\ w\left(\frac{i}{n}\right). \tag{3.6}$$

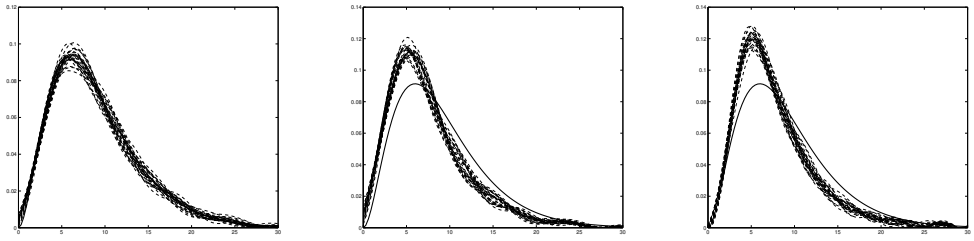
where $Z_{(i)}$ denotes the i-th order statistic associated with $(Z_1, \dots, Z_n)$ satisfying $Z_{(1)} \le \cdots \le Z_{(n)}$. We see that $\hat{L}$ is a linear combination of order statistics, also called an L-statistics in the literature.

3.2.2 State of the art

The pile-up and the biased data model are related to survival analysis. The former models some random right-censoring, while the latter is similarly defined as the nonlinear transformation model in [109], and it can also be viewed as a biased data problem with known bias as in [110].

The first problem that we studied concerns the pile-up model, where we consider additional measurement errors as defined in (3.2) to stick closely to the nature of fluorescence lifetime measurements. Combining nonlinear distortions with additive noise is new in the literature and real technical difficulties have to be faced in order to preserve standard deconvolution rates. We show that deconvolution methods in the spirit of [111], [112], [113] or [114], whose use in survival analysis is unusual, can be adapted to the pile-up model to derive oracle-type risk bounds.

Our second work adresses the problem of nonparametric density estimation in the general biased data model. It is noteworthy that the model can be related to other biased data contexts, which have been studied from various points of view by several authors: strategies for estimating cumulative distribution functions are proposed by [115], [116], [117], [118], [119]. Adaptive projection estimators correspond to methods originally described by [120] and applied to survival analysis and biased data by [121, 118] and [110]. We explore different strategies to invert the nonlinear relation of the target distribution and propose both kernel and projection estimators. For all estimators the question of adaptive model selection or bandwidth selection is addressed. Concerning the kernel estimators we follow the recent promising approach of Goldenshluger and



(a) (b) (c)

Figure 3.1: (a) Estimation with pile-up correction and deconvolution. (b) No pile-up correction. (c) No deconvolution.

Lepski [122]. An extensive simulation study compares all estimators and provides new insights on their performances.

3.2.3 Estimator in the pile-up model with measurement errors

In [CR12] we consider the pile-up model defined in Section 3.2.1 with additional noise, that is, where the $(X_{i,k})_{i,k\geq 1}$ have the form (3.2) and the density of X is assumed to be the convolution of densities f and f_η, i.e. $f_{Y+\eta} = f \star f_\eta$, it is natural to consider deconvolution techniques based on Fourier transforms. Recall that $f^* = f^*_{Y+\eta}/f^*_\eta$, and by the Fourier inverse formula

$$f(z) = \frac{1}{2\pi}\int \mathrm{e}^{izt}\frac{f^*_{Y+\eta}(t)}{f^*_\eta(t)}\mathrm{d}t. \tag{3.7}$$

To construct an estimate of f, first consider the estimation of $f^*_{Y+\eta}(t) = \mathbb{E}(\mathrm{e}^{-it(Y+\eta)})$. According to (3.6), a moment estimator is given by

$$\widehat{f^*_{Y+\eta}}(t) = \frac{1}{n}\sum_{k\in[\![n]\!]} w_\theta(k/n)\mathrm{e}^{-itZ_{(k)}},$$

with weight function $w_\theta(u) = (1-\mathrm{e}^{-\theta})/(\theta(1-u(1-\mathrm{e}^{-\theta})))$ that depends here on the Poisson parameter θ and defined more generally in (3.4). By plugging $\widehat{f^*_{Y+\eta}}$ into (3.7), and if the noise distribution f_η is known, we finally get an estimator of f by

$$\hat{f}_m(z) = \frac{1}{2\pi n}\sum_{k\in[\![n]\!]} w_\theta(k/n)\int_{-\pi m}^{\pi m}\frac{\mathrm{e}^{it(z-Z_{(k)})}}{f^*_\eta(t)}\mathrm{d}t, \tag{3.8}$$

where a cut-off in the integral, here at $-\pi m$ and πm, is required, since the Fourier transform $f^*_\eta(t)$ tends to 0 when $|t|\to\infty$. Figure 3.1 displays simulation results to demonstrate that both corrections, the pile-up correction via the weights $w_\theta(k/n)$, and the noise correction, via the deconvolution, are necessary to recover the target density.

Alternatively, $\hat{f}_m$ can be written as a weighted kernel deconvolution estimator. This allows to make a link with many other works in the kernel deconvolution setting, see [113], [111], [123].

Furthermore, the alternative expression of $\hat{f}_m$ as a kernel estimator is an L-statistics, and thus the central limit theorem for L-statistics proved in [RRS10] (see Appendix B therein) could be applied to obtain asymptotic normality of $\sqrt{n}(\hat{f}_m(z) - f_m(z))$ with computable limit variance.

From a theoretical point of view, we obtain the following bound of the mean integrated squared error (MISE) of $\hat{f}_m$.

Proposition 3. *Let f_m denote the function verifying $f_m^* = f^* \mathbb{1}_{[-\pi m, \pi m]}$. Then, under some regularity assumptions, there is a constant C such that*

$$\mathbb{E}(\|\hat{f}_m - f\|^2) \leq \|f - f_m\|^2 + C\frac{\Delta_\eta(m)}{n} \quad \textit{where} \quad \Delta_\eta(m) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m} \frac{\mathrm{d}u}{|f_\eta^*(u)|^2}.$$

There is an explicit expression of the constant C, which depends on the Poisson parameter θ. Indeed, when θ increases, the minimum in (3.1) is taken over more and more variables and so the pile-up distortion is getting stronger and the estimation problem more difficult. Intuitively, this goes in hand with a larger constant C and so with a larger risk bound.

The above risk bound has a classical bias-variance decomposition. The bias $\|f - f_m\|^2 = (2\pi)^{-1}\int_{|u|\geq \pi m} |f^*(u)|^2 \mathrm{d}u$ is clearly decreasing when m increases. On the contrary, the variance term $\Delta_\eta(m)/n$ is increasing with m. Hence, a good choice of m operates a bias-variance trade-off.

To determine the rate of convergence of the MISE, it is necessary to specify the type of the noise distribution, namely the rate of decrease to 0 of f_η^* near infinity, as the variance depends crucially on it. In the fluorescence setting, one can show that it is appropriate to consider noise that is *ordinary smooth* of order γ, denoted by $\eta \sim OS(\gamma)$. This means that the Fourier transform f_η^* satisfies

$$c_0(1+u^2)^{-\gamma} \leq |f_\eta^*(u)|^2 \leq C_0(1+u^2)^{-\gamma}.$$

In classical deconvolution the regularity spaces used for the functions to estimate are Sobolev spaces defined by

$$\mathcal{C}(a, L) = \left\{ g \in (\mathbb{L}^1 \cap \mathbb{L}^2)(\mathbb{R}),\ \int (1+u^2)^a |g^*(u)|^2 \mathrm{d}u \leq L \right\}.$$

The optimization of the upper bound of the MISE provides the optimal choice of m and we obtain the following optimal rate of convergence for the MISE.

Proposition 4. *If $f \in \mathcal{C}(a, L)$ and $\eta \sim OS(\gamma)$, then for $m_{\mathrm{opt}} = O(n^{1/(2a+2\gamma+1)})$ it holds*

$$\mathbb{E}(\|\hat{f}_{m_{\mathrm{opt}}} - f\|^2) = O(n^{-2a/(2a+2\gamma+1)}).$$

Obviously, in practice the optimal choice m_{opt} is not feasible since a and part of the constants involved in the order are unknown. Therefore, a data-driven model selection device is required to choose a relevant $\hat{f}_m$ in the collection. Another issue in practice is that the noise distribution f_η and the Poisson parameter θ are usually unknown. They may be estimated, but the question is how plug-in estimates of f_η and θ affect the performance. We address these three issues in the following.

3.2.4 Automatic cut-off selection

Data-driven model selection can be performed following the classical penalization approach by Barron *et al.* [120]. To this end a different view of the estimator, namely as a minimizer of a contrast, is useful. Indeed, a good estimator of f is the function h that minimizes the difference $\|h-f\|^2$ over a given set of functions. An empirical approximation (up to a constant) of the difference $\|h-f\|^2 = \|h\|^2 - 2\langle h, f\rangle + \|f\|^2$ is given by the contrast γ_n defined by

$$\gamma_n(h) = \|h\|^2 - \frac{1}{\pi}\int h^*(-u)\frac{\widehat{f^*_{Y+\eta}}(u)}{f^*_\eta(u)}\mathrm{d}u.$$

Here it is natural to consider the set of functions $S_m = \{h, \quad \text{support}(h^*) \subset [-\pi m, \pi m]\}$. Then one can show that $\hat{f}_m$ defined in (3.8) minimizes the contrast γ_n over S_m, that is, $\hat{f}_m = \arg\min_{h\in S_m}\gamma_n(h)$. This also means that $\hat{f}_m$ is indeed a projection estimator, which is, by the way, also an advantage for its numerical evaluation, as it can be written as a sum.

The general method to select the cut-off parameter or model m consists in finding a data-driven penalty pen(.) such that the model $\hat{m}$ defined as

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}\{\gamma_n(\hat{f}_m) + \text{pen}(m)\} \tag{3.9}$$

achieves a bias-variance trade-off, where $\mathcal{M}_n$ is the collection of considered models. As γ_n is an approximation of the squared bias, the penalty is usually chosen to have the same order as the variance term. It is difficult to derive the optimal constants for the penalty, so numerical constants κ and κ' are introduced, which have to be calibrated by simulation. Here, we propose two penalties that are proven to be convenient in different contexts.

Theorem 8. *Let*

$$\text{pen}_1(m) = \kappa_1\left(a_\theta + \kappa_1' b_\theta \log(n)\right)\frac{\Delta_\eta(m)}{n}, \qquad \text{pen}_2(m) = \kappa_2\left(a_\theta + \kappa_2' c_\theta\right)\frac{\Delta_\eta(m)}{n},$$

where $a_\theta, b_\theta, c_\theta$ are known constants and κ_j, $\kappa_j', j=1,2$ are numerical constants to be calibrated via simulations. Let $\hat{f}_{\hat{m}}$ be the estimate defined in (3.8) *with $\hat{m}$ chosen according to* (3.9) *with one of the penalties* pen_1 *or* pen_2*. Then there are constants C, C' such that*

$$\mathbb{E}\left(\|\hat{f}_{\hat{m}} - f\|^2\right) \le C\inf_{m\in\mathcal{M}_n}\left(\|f - f_m\|^2 + \text{pen}(m)\right) + C'\frac{\log(n)}{n}. \tag{3.10}$$

Risk bounds of the form (3.10) are called oracle inequality, since the data driven estimator $\hat{f}_{\hat{m}}$ achieves the bias-variance compromise, up to the multiplicative constant C and the residual $C'\frac{\log(n)}{n}$. Compared to classical deconvolution results, the penalty pen_1 contains an additional $\log(n)$-term and thus induces a loss with respect to the expected rate. That is, when considering Sobolev spaces, the MISE is of the order $O\left((n/\log(n))^{-2a/(2a+2\gamma+1)}\right)$. This loss is certainly due to the complexity of the problem under consideration, which involves several sources of error, namely a nonlinear distortion, additional noise and the estimation of the cdf G used in the weights.

To improve the bound and avoid the log-loss, we can use the second penalty pen_2. However, to prove the bound, we found it necessary to split the data sample into two parts, say $(Z_{-i})_{i\in[\![n]\!]}$ and $(Z_i)_{i\in[\![n]\!]}$ to separate the estimation of the cdf G from the rest. That is, an independent estimate $\tilde{G}_n(t) = \frac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{1}\{Z_{-i} \le t\}$ of G is computed and used in the weights $w_\theta(\tilde{G}_n(Z_i))$ evaluated at the observations $(Z_i)_{i\in[\![n]\!]}$ of the second sample. More precisely, the weights $w_\theta(k/n) = w_\theta(\hat{G}_n(Z_{(k)}))$ in (3.8) are replaced with $w_\theta(\tilde{G}_n(Z_i))$. Then the optimal rate of convergence is obtained, that is, the risk bound is of order $O\left(n^{-2a/(2a+2\gamma+1)}\right)$ according to the above theorem.

3.2.5 Practice-oriented setting

In time-resolved fluorescence, the Poisson parameter θ of the number N of photons hitting the detector is unknown, but easily estimated. As there are excitations that are not followed by the emission of any photon, N has indeed not a restricted, but a classical Poisson distribution with values in $\{0, 1, \dots\}$. So we can use the proportion P_0 of excitations that are not followed by the observation of a photon to estimate θ. As $\mathbb{P}(N = 0) = \mathrm{e}^{-\theta}$, a natural estimate of θ is given by

$$\hat{\theta} = -\log(P_0).$$

For the case where the estimate $\hat{\theta}$ is used instead of θ in the definition of $\hat{f}_m$ in (3.8), we propose to use a new penalty, which is now a random quantity, inducing new difficulties in the proof. The penalty is conceived such that the above risk bound still holds (with a slight loss in the constants) and the bound has the same order of magnitude as before. This means that the estimation procedure is robust with respect to this additional estimation step.

Theorem 9. *Let*

$$\widehat{\mathrm{pen}}(m, \hat{\theta}) = \kappa_3 \left(a_{\hat{\theta}} + \kappa_3'(b_{\hat{\theta}} + d)\log(n)\right) \frac{\Delta_\eta(m)}{n}.$$

Let $\hat{f}_{\hat{m}}$ be the estimate defined in (3.8) *using $\hat{\theta}$ instead of θ and with $\hat{m}$ chosen according to* (3.9) *with the penalty $\widehat{\mathrm{pen}}(m, \hat{\theta})$. Then the risk bound given in* (3.10) *holds with appropriate constants C, C'.*

Another issue in practice concerns the noise distribution f_η, which may not be the known and may be of nonparametric form. In the fluorescence set-up, a large independent sample of pure noise, say $(\eta_{-k})_{k\in[\![m]\!]}$, may be available, which can be used to estimate f_η^* by $\hat{f}_\eta^*(u) = \frac{1}{m}\sum_{k\in[\![m]\!]} \mathrm{e}^{-iu\eta_{-k}}$ and replaced in our procedure. In [124] the same substitution is considered for deconvolution methods and it is shown that for ordinary smooth noise and large sample sizes this leads to a risk bound exactly analogous to the one given in (3.10). In a numerical study we showed that there is nearly no loss when using an estimated $\hat{f}_\eta^*$ instead of the exact f_η^*. A rigorous theoretical justification would clearly require a considerable amount of work due to measurement errors and the nonlinear distortion.

3.2.6 Nonparametric weighted estimators for biased data

Now in our second study in [CR16] we consider the general biased data model defined in (3.3) and the problem of constructing a nonparametric estimate of the underlying target density f. Indeed, the two standard approaches in nonparametric estimation, namely kernel and projection estimation, can be applied here. Furthermore, there are two ways to correct the bias: either a classical density estimate is computed directly on the data and then a correction is applied, as in [125], or weights are directly associated with the data so that a direct estimator of the quantity of interest is obtained, as in [RRS10] and in the previous pile-up model with measurement errors. In both cases, adaptive devices for the selection of the kernel bandwidth or the model can be established. It is most interesting to compare all these estimators, which is done in this work.

To start with, let K be a kernel, h a bandwidth and $K_h(u) = K(u/h)/h$. The standard kernel estimator of g based on observations $Z_1, \dots, Z_n$ from the distribution G is given by

$$\hat{g}_h^{\text{ker}}(x) = \frac{1}{n} \sum_{i \in [\![n]\!]} K_h(x - Z_i), \quad x \in \mathbb{R}.$$

Then, using to (3.4), a *plug-in estimator* of f is given by

$$\hat{f}_h^{\text{ker-P}}(x) = w(\hat{G}_n(x)) \hat{g}_h^{\text{ker}}(x), \quad x \in \mathbb{R}.$$

Alternatively, taking $\psi(z) = K_h(x - z)$ in (3.5), according to (3.6) another estimator is given by

$$\hat{f}_h^{\text{ker-W}}(x) = \frac{1}{n} \sum_{i \in [\![n]\!]} w(\hat{G}_n(Z_i)) K_h(x - Z_i).$$

Concerning projection estimators, the general idea is to approximate g (or f) by its orthogonal projection onto some function space. Let A be an interval and $(\varphi_j)_{j \geq 1}$ an orthonormal basis of $\mathbb{L}^2(A)$. Denote the subspaces $S_m = \text{Span}(\varphi_j, j \in [\![d_m]\!])$ of dimension d_m. The orthogonal projection of g on S_m is given by $g_m = \sum_{j \in [\![d_m]\!]} a_j \varphi_j$ with coefficients $a_j = \langle g, \varphi_j \rangle = \mathbb{E}_{Z \sim G}[\varphi_j(Z)]$, which can be estimated by $\hat{a}_j = n^{-1} \sum_{i \in [\![n]\!]} \varphi_j(Z_i)$. Hence, an estimate of g is given by $\hat{g}_m^{\text{proj}} = \sum_{j \in [\![d_m]\!]} \hat{a}_j \varphi_j$, and finally, an estimator of f is obtained by

$$\hat{f}_m^{\text{proj-P}}(x) = w(\hat{G}_n(x))\ \hat{g}_m^{\text{proj}}(x), \quad x \in \mathbb{R}.$$

To apply the second bias-correction method, the orthogonal projection of f on S_m is given by $f_m = \sum_{j \in [\![d_m]\!]} b_j \varphi_j$ with coefficients $b_j = \langle f, \varphi_j \rangle = \mathbb{E}_{Y \sim F}[\varphi_j(Y)]$. With $\psi = \varphi_j$ in (3.5), the coefficient b_j is approximated by $\hat{b}_j = n^{-1} \sum_{i \in [\![n]\!]} w(i/n) \varphi_j(Z_{(i)})$. Hence, a second projection-type estimator of f is given by

$$\hat{f}_m^{\text{proj-W}}(x) = \sum_{j \in [\![d_m]\!]} \hat{b}_j \varphi_j(x), \quad x \in \mathbb{R}.$$

In the following $(\varphi_j)_{j \geq 0}$ is the trigonometric basis and we consider subspaces S_m of dimension $d_m = 2m + 1$. This basis has the advantage of simplicity and provides nested models allowing

for fast computation.

For the four density estimators risk bounds of very similar form are obtained. One can consider both, the pointwise and the integrated risk. For instance, for the estimator $\hat{f}_h^{\text{ker-W}}$ we prove the following bounds.

Proposition 5. *Under appropriate smoothness assumptions,*
(i) for any x_0, there is a constant C_1 such that

$$\mathbb{E}\left[(\hat{f}_h^{\text{ker}-\text{W}}(x_0) - f(x_0))^2\right] \leq 3(K_h * f(x_0) - f(x_0))^2 + \frac{C_1\|f\|_\infty}{nh}. \tag{3.11}$$

(ii) there exists a constant C_2 such that satisfies

$$\mathbb{E}\left[\|\hat{f}_h^{\text{ker}-\text{W}} - f\|_2^2\right] \leq 3\|K_h * f - f\|^2 + \frac{C_2}{nh}.$$

As usual, the risk bounds are composed of a squared bias term and a variance term. The first one decreases when $h \to 0$, whereas the second increases. Hence, automatic bandwidth selection aims at finding a compromise between these two antagonist terms.

More precise orders for the bias term may be obtained under stronger assumptions on the regularity of the kernel and when the density f belongs to a Hölder or the Nikol'ski space. If the bandwidth h is chosen of order $n^{-1/(2\beta+1)}$, where β is the Hölder (or the Nikol'ski) regularity index, then the resulting rate of the MISE is of order $n^{-2\beta/(2\beta+1)}$. But as β is unknown, this choice cannot be done in that naive way and thus data-driven methods for bandwidth selection are required.

Similar results are obtained for the projection estimators, as, for instance, for $\hat{f}_m^{\text{proj-W}}$.

Proposition 6. *There is a constant C_3 such that*

$$\mathbb{E}\left[\|\hat{f}_m^{\text{proj}-\text{W}} - f\mathbb{1}_A\|_2^2\right] \leq \|f\mathbb{1}_A - f_m\|_2^2 + C_3\frac{d_m}{n}.$$

One can show that on Besov spaces $\mathcal{B}_{\alpha,2,\infty}(A)$, choosing $d_{m^*} = O(n^{1/(2\alpha+1)})$ yields that $\mathbb{E}(\|\hat{f}_{m^*} - f_A\|_2^2) = O(n^{-2\alpha/(2\alpha+1)})$. This rate is known to be optimal in the minimax sense for density estimation for direct observations [126].

3.2.7 Data-driven bandwidth selection

To develop devices for a data-driven selection of the bandwidth h, we follow the recent approach of Goldenshluger and Lepski [122] that relies on empirical processes and powerful deviation inequalities and offers convenient and rigorous control of the estimators. Part of the results obtained by this approach are nonasymptotic, contrary to many kernel studies. Applying this method in the case of biased data is a novelty, in theory and in practice.

We illustrate the construction of the selector of the best bandwidth for the estimator $\hat{f}_h^{\text{ker-W}}(x_0)$ of $f(x_0)$ at some fixed point x_0. Let $\mathcal{H}$ be a finite collection of bandwidths. Essentially, Goldenshluger and Lepski propose an improved estimation of the squared bias in the upper bound

of the MISE in (3.11). To this end, we first introduce new estimators $\hat{f}^{\text{ker-W}}_{h,h'}$ of f depending on two bandwidths h, h' defined as

$$\hat{f}^{\text{ker-W}}_{h,h'}(x) = K_{h'} * \hat{f}^{\text{ker-W}}_{h}(x), \quad x \in \mathbb{R}.$$

The specific idea here is that $K_{h'} * (K_h * f - f)$ is approximately $K_h * f - f$, and so for small h', $\left(\hat{f}^{\text{ker-W}}_{h,h'}(x_0) - \hat{f}^{\text{ker-W}}_{h'}(x_0)\right)^2$ is a good approximation of the squared bias. Unfortunately, this estimate has a bias itself, which is of the same order as the variance. Hence, the new estimator of the squared bias is defined by

$$B(h, x_0) = \sup_{h' \in \mathcal{H}} \left[\left(\hat{f}^{\text{ker-W}}_{h,h'}(x_0) - \hat{f}^{\text{ker-W}}_{h'}(x_0) \right)^2 - V(h') \right]_+ \quad \text{with} \quad V(h) = \kappa D \|f\|_\infty \frac{\log n}{nh},$$

where D is a known constant. The term $V(h)$ can be interpreted as a variance estimate, augmented by a $\log(n)$ factor. Consequently, the optimal bandwidth $h^{\text{ker-W}}(x_0)$ is given by

$$h^{\text{ker-W}}(x_0) = \arg\min_{h \in \mathcal{H}} \{B(h, x_0) + V(h)\}.$$

Here, the existence of a minimal value for the constant κ in the variance term $V(h)$ has been studied only very recently in [127], and the calibration procedures are not yet well understood. This is probably due to the fact that the variance estimate $V(h)$ plays two different roles, namely as a variance estimate and and as bias correction. This is why our numerical study of Section 3.2.8 is of interest.

Now the following result holds for the estimator $\hat{f}^{\text{ker-W}}_{\hat{h}^{\text{ker-W}}(x_0)}(x_0)$.

Theorem 10. *Under some regularity assumptions, there are constants C and $\bar{C}$ such that for all $\kappa \geq \kappa_{\min}$,*

$$\mathbb{E}\left[\left(\hat{f}^{\text{ker-W}}_{\hat{h}^{\text{ker-W}}(x_0)}(x_0) - f(x_0) \right)^2 \right] \leq C^* \inf_{h \in \mathcal{H}} \left(\|K_h * f - f\|^2_\infty + V(h) \right) + \bar{C} \frac{\log n}{n}.$$

If f belongs to the Hölder class $\Sigma(\beta, L)$ and if $\mathcal{H}$ is large enough, the upper bound is of order $(n/\log(n))^{-2\beta/(2\beta+1)}$. Moreover, in classical density estimation (without any nonlinear distortion), the $\log(n)$-loss is known to be unavoidable and thus adaptive minimax (see [128]).

Concerning the plug-in kernel estimator $\hat{f}^{\text{ker-P}}_h(x_0)$, one can proceed in the same way, namely by adapting the new bias estimator. This results in an optimal bandwidth of the form $h^{\text{ker-P}}(x_0) = \arg\min_{h \in \mathcal{H}} \left\{ \tilde{B}(h, x_0) + \tilde{V}(h) \right\}$ and a similar oracle inequality is derived. Under appropriate regularity assumptions, the risk bound on $\hat{f}^{\text{ker-P}}_{h^{\text{ker-P}}(x_0)}(x_0)$ is an automatic compromise related to the regularity of g and provides the best possible rate if f and g belong to the same Hölder space.

Global bandwidth selection for both kernel estimators follows the same lines, essentially by replacing squared differences by squared norms. Oracle bounds are of the same flavor.

Concerning model selection for the projection estimators the approach used in Section 3.2.4 can be adapted. That is why it is not presented in detail here.

From a theoretic point of view, all the procedures are proved to deliver the best possible

tradeoff when selecting the model or the bandwidth and have (nearly-)optimal rates. Now it is intrinsically interesting to assess their numerical performances.

3.2.8 Experimental study

In an extensive numerical study all adaptive estimators are compared on synthetic data from the pile-up model. Recall that there are two bias correction strategies both applicable to projection and kernel estimators. Furthermore, for each kernel estimator with a given bias correction both pointwise and global bandwidth selection can be used, resulting in a total of six estimators to be compared one to another. All numerical constants κ are calibrated via simulation. The projection estimators are rather robust to the choice of the value of κ, while for the kernel estimators calibration is much less evident. In the experimental study we make the following observations.

Bias-correction approach: weighted estimators or plug-in strategy? Often both strategies provide very similar results. But in cases of a strong pile-up effect, the weighted estimators are slightly doing better and are more robust than their plug-in counterparts.

Estimation approach: Global kernel, pointwise kernel or projection strategy? Projection and global kernel estimators give the best overall results. For some types of densities projection estimators are clearly the best. Interestingly, the pointwise kernel estimators are mostly far behind all other estimators. This is surprising as the pointwise method conceptually outplays the global one, since it is conceived to capture peaks like in the exponential or Laplace distribution. An analysis of the oracle (see next paragraph) sheds more light on this issue.

Comparison to the oracle Here the oracle is the MISE of the best estimator that could have been chosen and that can be evaluated numerically in simulations. More precisely, for instance, for the weighted kernel estimator $\hat{f}^{\text{ker-W}}_{\hat{h}^{\text{ker-W}}}$ with global bandwidth selection $\hat{h}^{\text{ker-W}}$, the oracle for a given dataset is $\min_{h\in\mathcal{H}} \|\hat{f}^{\text{ker-W}}_h - f\|^2$.

Our numerous simulations make clear that the pointwise kernel methods perform much better than their global counterparts, often a factor 4 between the different oracles (depending on the type of the underlying target distribution f). This is coherent with our understanding of the pointwise selection approach, where the optimal bandwidth is chosen at every point x_0, while the global method selects a single bandwidth that is used for the entire estimation interval. As our projection estimators also rely on a global selection method for the entire interval, it is natural that their oracles are much worse than those of the pointwise kernel methods.

Now it is interesting to analyze the difference of the oracle with the actually achieved MISE by the adaptive estimators. Obviously, the kernel pointwise estimator is not able to take advantage of its very small oracles, as there is a factor 10 to 20 between the oracles and the corresponding MISE values. It is evident that the pointwise bandwidth selection fails completely. For the global kernel estimators the loss between the oracle and the realized MISE by the estimator is only about a factor 2, and projection estimators do even better.

Chapter 4

Topics in machine learning

This chapter provides a collection of works on various topics in signal processing and machine learning. While they are not really connected one to another, they all aim at improving some statistical method to provide more relevant results in practice. We improve the estimation of radar signals by taking into account the complex data structure (Section 4.1), we reduce the dimension of high-dimensional data using recent results on random matrices (Section 4.2), we handle missing data in self-organizing maps (Section 4.3) and show how to avoid mistakes when clustering data points (Section 4.4).

4.1 Estimation of multipath radar signals

A transmitted, unknown radar signal is intercepted. Due to propagation and reflexion on obstacles, the signal arrives via more than one pathway with different time delays on the receiver, which composed of several sensors. The aim in signal intelligence (SIGINT) is to recover not only the waveform of the signal, but also the direction of arrival. Exploiting the parsimonious time-frequency representation of the signal, we write the model as a linear model with structured sparsity pattern and propose a new orthogonal matching pursuit algorithm for the inference that is suitable for large dimensions. Our method performs well even when the signal-to-noise ratio is low.

This work is the result of my postdoc at Télécom ParisTech on a project with Direction générale de l'Armement (DGA). It is joint work with Maurice Charbit and Céline Lévy-Leduc available in [RLLC11a, RLLC11b].

4.1.1 State of the art

Several subspace methods and maximum likelihood approaches have been proposed to deal with coherent sources, see [129, 130, 131, 132]. However, all of them are parametric approaches and to the best of our knowledge, estimating the waveform, which is an infinite-dimensional parameter, in a nonparametric way has not yet been considered for multipath signals.

Our approach consists in a linearization of the model yielding a high-dimensional linear model with sparse coefficients. For dealing with the estimation in sparse linear regression models, the

Lasso [133] and the greedy orthogonal matching pursuit (OMP) ([134], [135]) have become very popular tools. In our case, an inspection of the model shows that the parameter vector is not sparse in an arbitrary way, but sparsity is subject to a number of constraints so that the allowed sparsity patterns have specific structure. In order to include prior information on the sparsity structure, different approaches have been proposed in the literature. On the one hand, there are methods based on composed ℓ_1/ℓ_2-penalties (elastic nets [136]; fused Lasso [137]; group Lasso [138]; composite absolute penalty [139]; overlapping groups [140, 141]). On the other hand, an extension of the orthogonal matching pursuit for structured solutions has been proposed in [142] that applies to non-overlapping groups. Compared to Lasso methods, OMP algorithms have the advantage to be more easily scalable to high dimension, which is of most interest in our case. We follow the line of research for OMPs and propose an extension to consider more sophisticated sparsity structures, namely overlapping and nested groups.

4.1.2 Modelling multipath radar signals

Let $\{s_0(t)\}_{t\geq 0} \subset \mathbb{C}$ be the original radar signal emitted by the source, where t is the time. The number U of propagation pathways is unknown. Every pathway is characterized by a direction of arrival d_u, a time-delay t_u and an attenuation constant a_u. The signal is detected by an antenna with C sensors and its array response is a known function $\mathbf{r}$ taking values in $\mathbb{R}^C$. The signal arriving at the sensor array is given by

$$\mathbf{s}(t) = \sum_{u\in[\![U]\!]} a_u s_0(t-t_u)\mathbf{r}(d_u).$$

As the time lags t_u are much shorter than the length of the emitted signal s_0, the arriving $\mathbf{s}$ is the superposition of several delayed replicates. The signal is observed at time steps of length Δ, so that observations are of the form $\mathbf{y}_m = \mathbf{s}(m\Delta) + \boldsymbol{\varepsilon}_m$, $m \in [\![M]\!]$, where $\boldsymbol{\varepsilon}_m$ denotes additive Gaussian noise.

4.1.3 Sparse linear model with structured sparsity pattern

As the relationships between the parameters are very complex, we propose a reformulation by linearizing the model. On the one hand, we represent the waveform in an overcomplete basis and, on the other hand, we discretize the other parameter spaces. This leads to a linear model with a high-dimensional model parameter and nonlinear constraints. More precisely, using a dictionary $\mathcal{D} = \{\varphi_j, j \in [\![J]\!]\}$ of waveforms, we assume that there are coefficients β_j such that the signal verifies

$$s_0 = \sum_{j\in[\![J]\!]} \beta_j \varphi_j.$$

Using grids $\{\tau_1, \dots, \tau_P\}$ and $\{\theta_1, \dots, \theta_Q\}$ of potential values for the delay times t_u and the angles of arrival d_u, and denoting $\alpha_{p,q} = \sum_{u\in[\![U]\!]} a_u \mathbb{1}\{t_u = \tau_p, d_u = \theta_q\}$, the observation $\mathbf{y}_m$ reads

$$\mathbf{y}_m = \sum_{j\in[\![J]\!]} \sum_{p\in[\![P]\!]} \sum_{q\in[\![Q]\!]} \alpha_{p,q}\beta_j \mathbf{r}(\theta_q)\varphi_j(m\Delta - \tau_p) + \boldsymbol{\varepsilon}_m, \qquad m \in [\![M]\!],$$

where most of the coefficients $\alpha_{p,q}$ and β_j are zero. Now, by storing all known quantities $\mathbf{r}(\theta_q)\varphi_j(m\Delta - \tau_p)$ in a matrix $\mathbf{X}$ and all model parameters $\alpha_{p,q}\beta_j$ in a vector w, the model can be written as a linear regression of the form

$$Y = \mathbf{X}w + \boldsymbol{\varepsilon}, \quad \text{with } w \in \mathcal{W} = \left\{ w = \beta \otimes \alpha \in \mathbb{C}^{JPQ} : \alpha \in \mathbb{C}^{PQ}, \beta \in \mathbb{C}^{J} \right\},$$

where $w = \beta \otimes \alpha$ denotes the Kronecker product. The set $\mathcal{W}$ is not the entire space $\mathbb{C}^{JPQ}$ but a smaller, nonconvex subset and the sparsity pattern $\mathcal{S} = \{(j,p,q) : w_{j,p,q} = 0\}$ of w is subject to a number of constraints. In fact, $\alpha_{p',q'} = 0$ implies that $w_{j,p',q'} = 0$ for all j. Likewise, if $\beta_{j'} = 0$ then $w_{j',p,q} = 0$ for all p and q. Thus, the sparsity pattern of w has a specifique structure.

4.1.4 Constraint relaxation

Clearly, for this model a method is required that works for high-dimensional feature spaces, as there are much more parameters than observations ($PQJ \gg MC$). Furthermore, the method has to account for the sparsity and the specific structure of the sparsity pattern of the model parameter. We propose a procedure in two steps, where the first step is based on a relaxation of the constraints on the model parameter, such that a regularization method can be applied, which has a penalty that induces structured sparsity similar to [139]. More precisely, solve the penalized minimization problem given by

$$\min_{w \in \mathbb{C}^{JPQ}} \left\{ \|Y - \mathbf{X}w\|^2 + \Omega(w) \right\}, \tag{4.1}$$

where $\Omega(w)$ is a structured ℓ_1/ℓ_2-penalty encoding the sparsity pattern of w of the form

$$\Omega(w) = \lambda_1 \sum_{p \in [\![P]\!]} \sum_{q \in [\![Q]\!]} \|w_{G^{\alpha}_{p,q}}\|_2 + \lambda_2 \sum_{j \in [\![J]\!]} \|w_{G^{\beta}_{j}}\|_2,$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters and $G^{\alpha}_{p,q}$ and G^{β}_{j} are the sets of indices indicating zero entries in w coming from $\alpha_{p,q} = 0$ or $\beta_j = 0$, respectively. The solution $\tilde{w}$ is a vector with admissible sparsity structure. However, there may not exist any vectors α and β such that $\tilde{w}$ equals the Kronecker product $\alpha \otimes \beta$. That is, this step only serves to estimate the sparsity pattern. Then, from the sparsity pattern one can derive the set of indices $\mathcal{I}_\alpha$ and $\mathcal{I}_\beta$ of coefficients $\tilde{\alpha}_{p,q}$ and $\tilde{\beta}_j$ that must be nonzero.

Then, in the second step of our procedure, the goal is to compute the best nonzero entries of such vectors $\tilde{\alpha}$ and $\tilde{\beta}$. To achieve this, the matrix $\mathbf{X}$ is reduced by keeping only the predictors $x_{j,p,q}$ such $(p,q) \in \mathcal{I}_\alpha$ and $j \in \mathcal{I}_\beta$. This reduces the dimension largely, and when $\tilde{\beta}$ is fixed, the model is linear in $\tilde{\alpha}$ with explicit ordinary least squares estimator. Thus, the Nelder-Mead simplex method can be used to compute the least squares estimator in the model

$$Y = \mathbf{X}^{\text{reduced}}(\tilde{\beta}^{\text{reduced}} \otimes \tilde{\alpha}^{\text{reduced}}) + \boldsymbol{\varepsilon},$$

where no constraints are put on $\tilde{\alpha}^{\text{reduced}}$ and $\tilde{\beta}^{\text{reduced}}$.

4.1.5 Scalability by orthogonal matching pursuit

To obtain a good representation of the waveform, a large dictionary shall be used. Likewise, to avoid biased estimators of angles and time delays, we may use fine grids. However, this results in a huge design matrix $\mathbf{X}$ (in our simulations we easily achieve more than 10^9 columns). This raises computational difficulties, namely concerning the storage of the design matrix $\mathbf{X}$. These considerations have motivated us to use an orthogonal matching pursuit algorithm to solve (4.1), resulting in a scalable algorithm.

Orthogonal matching pursuit (OMP) consists in adding iteratively the predictor to the current solution $w^{(t)}$ which is the most correlated with the current residual $r^{(t)} = Y - \mathbf{X}w^{(t)}$. In [142] OMP is extended to selecting non-overlapping groups of variables, where $\mathcal{G}$ is a set of pairwise disjoint sets G_g and one adds all variables of group G_{g^*} if $\|\mathbf{X}_{G_{g^*}}^T r^{(t)}\|_2^2 = \max_{g\in\mathcal{I}} \|\mathbf{X}_{G_g}^T r^{(t)}\|_2^2$. Proceeding in this way guarantees that at every step of the algorithm, the current solution respects the required sparsity structure. As our sparsity patterns are more involved with overlapping groups, the selection of the components to be activated in the current solution must be chosen more carefully. In [RLLC11a] we provide full technical details on this adapted choice of components.

The regularized problem (4.1) is similar to the active set algorithm described in [141], but our algorithm is faster and scalable to very large dimension. Indeed, where the active set algorithm solves the penalized problem of reduced dimension by a second-order cone programming, our updates of $w^{(t)}$ are performed by the ordinary least squares estimator, which is known explicitly.

Furthermore, in view of scalability it is important to note that the OMP-type algorithm does not perform any computations involving the entire design matrix. Essentially, only correlations of the current residual $r^{(t)}$ and the predictors $\mathbf{X}_G$ associated with a group of variables G are computed. That is, only a part of the matrix $\mathbf{X}$ is required. In short, the storage of $\mathbf{X}$ can be avoided by recomputing the required predictors at every iteration. Thus, there are almost no limits on the size of the regression matrix, and almost arbitrarily large dictionaries and grids may be used.

A simulation study illustrates the good performance of the new method and exhibits a considerable improvement with respect to some elementary method. Even when the signal to noise ratio is low, very accurate results are achieved.

4.2 Dimension reduction with random matrix theory

High-dimensional noisy data often live in a subsapce of low dimension. Dimensionality reduction aims at separating signal from noise in order to preserve significant properties of the data in a low-dimensional space before analyzing them by further statistical methods. We address the challenge of estimating the dimension of the subspace where the signal lives in and propose a novel estimator that relies on recent results from random matrix theory. Consistency of the estimator is proved in the modern asymptotic regime, where the number of features grows proportionally with the sample size. Experimental results show that the novel estimator is robust to noise and, moreover, it gives highly accurate results in settings where alternative methods

fail. This is joint work with Malika Kharouf and Nataliya Sokolovska, which is published in [KRS18].

4.2.1 State of the art

In many applications as speech recognition [143], wireless communications [144], hyperspectral imaging [145], chemometrics [146], medical imaging [147], genomics [148] or mathematical finance [149], the signal space dimension is much lower than the number of observed features. A challenge is to determine the low-dimensional signal space, in order to perform dimension reduction by projecting the data onto a the smaller subspace. A major difficulty in real data sets is the presence of noise, making the estimation of the signal space involved. Here we address the fundamental question of determining the optimal dimension of a high-dimensional problem.

An overview of methods to estimate the dimension of the signal space is provided in [150]. The most prominent method is based on the number of principal components that is necessary to explain a given part of the total variance [151]. The scree graph, which is the plot of ordered sample eigenvalues, is also widely used in practice. These selection methods are rather heuristic. A recent approach based on eigengaps, that is the distance between consecutive sample eigenvalues (for both white [152] and colored [153] noise), comes with theoretical guarantees stemming from results in random matrix theory. However, in practice, the eigengap method may provide erroneous results when the signal eigenvalues are not well separated.

The purpose of this work is to improve on the eigengap method to obtain robustness in the presence of rather close signal eigenvalues, while preserving the strong theoretical properties of the initial approach. This is achieved by a more global look on the sample eigenvalues.

4.2.2 Spiked population model

In the additive noise model the observed vector $Y \in \mathbb{R}^p$ equals a signal vector S corrupted by additive white noise E, that is,

$$Y = S + E,$$

where S and E are independent, and E is centered with covariance $\sigma^2 \mathrm{I}_p$. The covariance of Y verifies $\mathbf{Cov}(Y) = \mathbf{Cov}(S) + \sigma^2 \mathrm{I}_p$.

Often the signal S is a linear combination of a relatively small number of predictors, that is, $S = \mathbf{B}x$ for some $(p \times r)$-matrix $\mathbf{B}$ with $r < p$. In other words, the signal lives in a proper subspace of $\mathbb{R}^p$ of dimension r. To separate signal from noise, data may be compressed to this smaller subspace. This model is also referred to as the *spiked population model*, and the eigenvalues of the covariance matrix of the observed vector $\mathbf{Cov}(Y)$ denoted by $\lambda_1 \geq \cdots \geq \lambda_p > 0$ verify

$$\lambda_\ell = \begin{cases} \alpha_\ell + \sigma^2, & l \in [\![r]\!] \\ \sigma^2, & l > r \end{cases},$$

where $\alpha_1 > \cdots > \alpha_r > 0$ are the non zero eigenvalues of the signal's covariance $\mathbf{Cov}(S)$. The first r eigenvalues $\lambda_1, \ldots, \lambda_r$ are called *spikes* and they yield important information on the signal dimension r. Now consider a data set $(Y_1, \ldots Y_n)$ of n independent realizations of Y and

denote by $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ the eigenvalues associated with the sample covariance matrix of $(Y_1, \ldots, Y_n)$.

4.2.3 Eigenrange method

We propose an estimator that relies on asymptotic results on the eigenvalues. In the pure noise case, where $Y = E$ and $\mathbf{Cov}(Y) = \sigma^2 \mathrm{I}_p$, the seminal work of Marchenko and Pastur [154] shows that, when $p/n \to c$, the limits of all sample eigenvalues $\hat{\lambda}_\ell$ lie in the interval $[a, b]$ with $a = \sigma^2(1 - \sqrt{c})^2$ and $b = \sigma^2(1 + \sqrt{c})^2$. In the additive noise model, the nonspiked sample eigenvalues still tend to lie in the Marchenko-Pastur interval $[a, b]$, while the limits of the spikes are outside [155]. More formally,

$$\hat{\lambda}_\ell \longrightarrow \alpha_\ell + \sigma^2 \left(1 + c + \frac{c\sigma^2}{\alpha_\ell}\right) \quad a.s, \qquad \text{as } p/n \to c, \qquad \text{for } \ell \in [\![r]\!].$$

We see that sample eigenvalues are asymptotically biased. Furthermore, the first and last pure noise eigenvalues tend to the limits of the interval $[a, b]$. As a consequence, the range of the pure noise sample eigenvalues, that is, $\hat{\lambda}_{r+1} - \hat{\lambda}_m$ with $m = \min(n, p)$, tends to $b - a$, whereas the distance $\hat{\lambda}_l - \hat{\lambda}_m$ for any $\ell \in [\![r]\!]$ is significantly larger.

From this viewpoint, a natural estimator of the signal dimension r is derived from the number of sample eigenvalues contained in an interval of approximate length $b-a$. For a threshold $\varepsilon_n > 0$, an estimator of the signal space dimension, called the *eigenrange method*, is defined by

$$\hat{r}^{\text{range}} = \min\{\ell : \hat{\lambda}_{\ell+1} - \hat{\lambda}_m - (b - a) < \varepsilon_n\}.$$

4.2.4 Consistency result

We establish consistency of $\hat{r}^{\text{range}}$ in the modern asymptotic regime, when both the sample size and the number of features tend to infinity. This is most relevant for applications where the number of features p is of the order of the sample size n or larger. Generally, results in this regime provide better approximations of the finite sample situation than those obtained in the traditional regime, where the number of features p is fixed.

Theorem 11. *Let $(\varepsilon_n)_{n\geq 1}$ be such that $\varepsilon_n \to 0$ and $n^{2/3}\varepsilon_n \to \infty$ as $n \to \infty$. Then, under appropriate model assumptions,*

$$\hat{r}_n^{\text{range}} \longrightarrow r \quad a.s. \qquad \text{as } p/n \to c.$$

It is noteworthy that consistency is obtained without strong distributional assumptions like normality as it is the case of maximum-likelihood approaches and others.

The Marchenko-Pastur interval $[a, b]$ depends on the unknown variance σ. As σ can be estimated from the pure noise sample eigenvalues, we propose to alternate the estimation of r and the estimation of σ. In general, convergence is attained in very short time.

Table 4.1: Success rates of eigenrange (ER), eigengap (EG), kink (K) and PCA for explaining 80% of the variance on 1000 simulated datasets.

(a) distinct spikes, $r = 3$, $c = 3$

n	eigenrange	eigengap	kink	PCA
10	0.48	0.42	0.49	0.53
100	1.00	1.00	1.00	0.00
1000	1.00	1.00	1.00	0.00

(b) close spikes, $r = 5$, $c = 0.5$

n	eigenrange	eigengap	kink	PCA
50	0.62	0.00	0.51	0.00
500	1.00	0.14	0.77	0.00
5000	1.00	0.48	0.48	0.00

4.2.5 Robustness of the eigenrange method

In a numerical study that compares different methods to select the dimension, the method based on the number of principal components to explain a given part of the total variance fails completely. In contrast, the heuristic kink method that searches an elbow in the scree graph, the eigengap method and our eigenrange method give excellent results in settings, where spiked eigenvalues are well separated (Table 4.1 (a)). However, when the setting contains rather close spiked eigenvalues, then the eigenrange method clearly outperforms the alternative methods (Table 4.1 (b)). Indeed, methods that focus on local features of the scree graph underestimate the dimension r, while the global approach of the eigenrange method works correctly. In this sense the eigenrange method is robust to the presence of very close spiked eigenvalues, whereas alternative methods fail.

4.2.6 Application to classification

To illustrate the importance of the accurate estimation of the signal space dimension, we consider the classification task. We use data from the UCI machine learning repository, namely four life science data sets. First, we apply alternatively the kink, eigengap and eigenrange methods to estimate the dimension of the data, then several PCA variants are applied to reduce the dimension of the data. Finally, classification is performed by a support vector machine. The PCA variants considered here for projecting the data are the structured sparse PCA method by Jenattion *et al.* [156], a sparse PCA approach by Zou *et al.* [157], and the inverse power method applied to sparse PCA by Hein *et al.* [158]. A different approach by Mestre *et al.* [159] relies on new consistent estimators of the eigenvectors when $p/n \to c < 1$. Figure 4.1 displays the 10-fold cross validation test accuracy, and we see that the eigenrange method achieves the optimal performance on all tested data sets and generally provides better results than the kink or eigengap approaches.

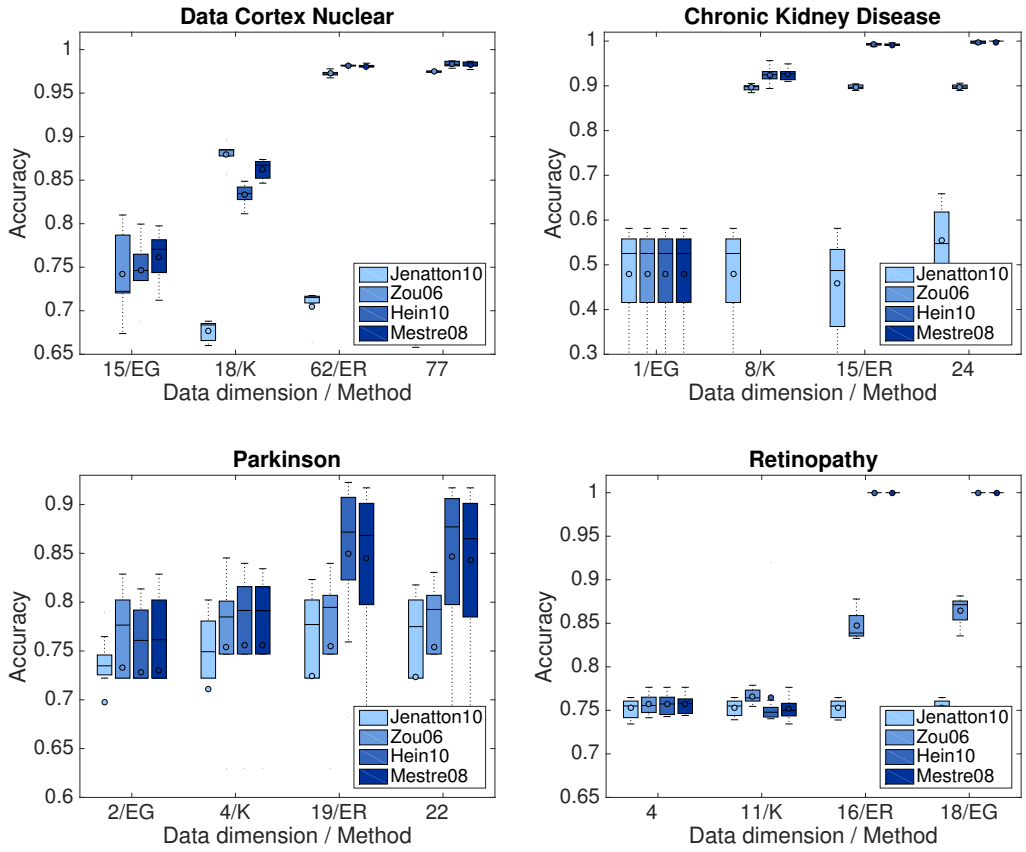


Figure 4.1: Classification accuracy on four life science data sets. Accuracy as a function of the estimated data dimension $\hat{r}$ (K – kink, ER – eigenrange, EG – eigengap) and different dimensionality reduction methods.

4.3 Self-organizing maps for incomplete data

A self-organizing map (SOM) is an unsupervised neural network which is widely used as a data exploration tool for data visualisation and clustering. The standard method is suitable only for complete data without any missing values. However, in many applications, partially observed data are the norm. With my student Sara Rejeb and in collaboration with Catherine Duveau at Safran Aircraft Engines, we propose an extension of SOM for incomplete data that incorporates the estimation of missing values. There is a published paper [RDR22] and an R package named `missSOM` [RDR22].

4.3.1 State of the art

Among the main tasks of data exploration are visualization and clustering of the data. While there is plethora of methods addressing one of the two tasks, self-organizing maps, introduced by [160], provide both a low-dimensional visual representation of the data in form of a map *and* a clustering of the observations. Self-organizing maps have become very popular in many fields of application, since they provide easily interpretable results with a global view of the data.

In practice, a common issue with datasets are missing entries, as they have a serious impact

on statistical results. They may lead to biased estimates and less accuracy. A first attempt to adapt SOM to partially observed presented in [161] consists in simply restricting all vector calculations to the observed entries. As such, there is no loss of information, since all observed data entries are taken into account in the algorithm. However, the method is not an imputation method.

In our work we propose a method that combines the learning of the map with the task of imputing missing values by a principled approach. Our motivation is the fact that any non trivial imputation method is based on some data model, and so it is natural to use the self-organizing map for imputation. Conversely, a better map may be learned when data are complete. Thus, treating both tasks simultaneously may be beneficial for the two of them.

4.3.2 The missSOM algorithm

To present our self-organizing map (SOM) for partially observed data, we first recall the standard method for complete data. A self-organizing map is a nonlinear projection of the high-dimensional data set, say $x_1, \ldots, x_n \in \mathbb{R}^p$, onto a two-dimensional map represented by a regular grid composed of K fixed neurons. The fixed spatial arrangement of the neurons on the map is the key for the preservation of the topology of the input data when projected onto the map. Every neuron k is associated with a p-dimensional prototype vector w_k, also called code vector, that is to be learned. The prototype vectors define a discretization of the data space, and each observation x_i is assigned to its closest prototype.

The Kohonen algorithm for standard SOM is an iterative procedure treating one randomly picked observation x_i per iteration. First, determine the neuron ℓ_{x_i} which is closest to x_i, that is, $\ell_{x_i} = \arg\min_{k \in [\![K]\!]} \|x - w_k\|_2$. Then, all code vectors w_k are updated by attracting them towards the measurement x_i. The attraction is the strongest for the winning neuron and weaker for distant neurons. More precisely, let $V_\lambda : [\![K]\!]^2 \mapsto \mathbb{R}_+$ be a neighborhood function defining the arrangement of the neurons on the map. Then, the update of the codes vectors is given by

$$w_k^{(t+1)} = w_k^{(t)} + \varepsilon_t V_\lambda(k, \ell_{x_i})(x_i - w_k^{(t)}), \quad k \in [\![K]\!], \tag{4.2}$$

for a sequence of learning steps $(\varepsilon_t)_{t \geq 0}$. Those updates eventually result in an ordered map, where neighboring neurons have similar prototype vectors.

Now, in the presence of missing entries, we propose to modfiy the Kohonen algorithm such that both the map and missing entries are learned simultaneously. With some abuse of notation, for a vector x_i, we denote its observed and missing parts by $x_i^{\text{complete}} = (x_i^{\text{obs}}, x_i^{\text{miss}})$. At iteration t, the current code vectors and imputed missing values obtained at the previous iteration are denoted by $w_k^{(t-1)}$ and $\hat{x}_i^{\text{miss}(t-1)}$, respectively. In our algorithm, the winning neuron is computed by restricting the Euclidean distance to the observed entries x_i^{obs}, that is, $\ell_{x_i^{\text{obs}}}^{(t)} = \arg\min_{k \in [\![K]\!]} \|x_i^{\text{obs}} - w_k^{\text{obs}(t-1)}\|_2$. The update of the code vectors is done according to (4.2), where missing values are imputed by the current values $\hat{x}_i^{\text{miss}(t-1)}$. Finally, we add a new step to update the imputed values $\hat{x}_i^{\text{miss}(t)}$ using a weighted mean of the code vectors given

Algorithm 2 Accelerated missSOM algorithm

Input: Data $x_1, \ldots, x_n$.
Initialization: Choose imputed values $\hat{\mathbf{x}}^{\text{miss}(0)}$ and code vectors $\mathbf{w}^{(0)}$.
Set $t = 1$.
while not converged **do**
  Set $\tilde{w}_i = w_i^{(t-1)}$ for $i \in [\![n]\!]$.
  for $i \in [\![n]\!]$ **do**
    Compute winning neuron $\ell_{x_i^{\text{obs}}}^{(t)} = \arg\min_{k \in [\![K]\!]} \|x_i^{\text{obs}} - w_k^{\text{obs}(t)}\|_2$.
    Update code vectors:
    for $k = [\![K]\!]$ **do**
      $w_k^{(t)} = \tilde{w}_k + \varepsilon_t V_{\lambda_t}(k, \ell_{x_i^{\text{obs}}}^{(t)}) \left((x_i^{\text{obs}}, \hat{x}_i^{\text{miss}(t-1)}) - \tilde{w}_k \right)$.
    end for
  end for
  Update imputed values according to (4.3).
  Set $t = t + 1$.
end while
Output: Final code vectors $\mathbf{w}^{(t)}$ and imputed data $\hat{\mathbf{x}}^{\text{miss}(t)}$.

by

$$\hat{x}_{i,j}^{\text{miss}(t)} = \frac{\sum_{k \in [\![K]\!]} V_{\lambda_t}(k, \ell_{x_i^{\text{obs}}}^{(t)}) w_{k,j}^{(t)}}{\sum_{k \in [\![K]\!]} V_{\lambda_t}(k, \ell_{x_i^{\text{obs}}}^{(t)})}. \tag{4.3}$$

The algorithm is summarized in Algorithm 2. Interestingly, the computing time of missSOM is comparable to the computing time of the Kohonen algorithm for complete data, as only the update of the imputed values is added, which is fast.

4.3.3 Loss function including imputed values

One can show that our approach has some theoretical foundation in the sense that the algorithm tends to minimize a new loss criterion. It is known from [162] that the classical Kohonen algorithm for complete data is a stochastic approximation algorithm for the minimization of the loss given by

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n} \sum_{i \in [\![n]\!]} \sum_{k \in [\![K]\!]} V_\lambda(k, \ell_{x_i}) \|x_i - w_k\|_2^2, \tag{4.4}$$

where $\mathbf{w} = (w_1, \ldots, w_K)$ are the code vectors. Now, for incomplete data, define a new loss by

$$\mathcal{L}_{\text{missom}}(\mathbf{w}, \hat{\mathbf{x}}^{\text{miss}}) = \frac{1}{n} \sum_{i \in [\![n]\!]} \sum_{k \in [\![K]\!]} V_\lambda \left(k, \ell_{x_i^{\text{obs}}} \right) \left\| (x_i^{\text{obs}}, \hat{x}_i^{\text{miss}}) - w_k \right\|_2^2,$$

where $(x_i^{\text{obs}}, \hat{x}_i^{\text{miss}})$ denotes the i-th observation completed with $\hat{x}_i^{\text{miss}}$ and $\hat{\mathbf{x}}^{\text{miss}} = (\hat{x}_1^{\text{miss}}, \ldots, \hat{x}_n^{\text{miss}})$. The criterion $\mathcal{L}_{\text{missom}}$ can be decomposed into two parts as

$$\mathcal{L}_{\text{missom}}(\mathbf{w}, \hat{\mathbf{x}}^{\text{miss}}) = \mathcal{L}_{\text{obs}}(\mathbf{w}) + \mathcal{L}_{\text{miss}}(\mathbf{w}, \hat{\mathbf{x}}^{\text{miss}}),$$

where $\mathcal{L}_{\text{obs}}$ is the counterpart of $\mathcal{L}$ in (4.4) in the presence of missing data and $\mathcal{L}_{\text{miss}}(\mathbf{w}, \hat{\mathbf{x}}^{\text{miss}})$ is given by

$$\mathcal{L}_{\text{miss}}(\mathbf{w}, \hat{\mathbf{x}}^{\text{miss}}) = \frac{1}{n} \sum_{i \in [\![n]\!]} \sum_{k \in [\![K]\!]} V_\lambda \left(k, \ell_{x_i^{\text{obs}}}\right) \left\| \hat{x}_i^{\text{miss}} - w_k^{\text{miss}} \right\|^2 .$$

A natural algorithm to minimize $\mathcal{L}_{\text{missom}}(\mathbf{w}, \hat{\mathbf{x}}^{\text{miss}})$ consists in alternating (i) the minimization of $\mathbf{w} \mapsto \mathcal{L}_{\text{obs}}(\mathbf{w})$, and (ii) the minimization of $\hat{\mathbf{x}}^{\text{miss}} \mapsto \mathcal{L}_{\text{miss}}(\mathbf{w}, \hat{\mathbf{x}}^{\text{miss}})$ with fixed $\mathbf{w}$. Here, (i) is roughly the classical Kohonen algorithm with the suitable way to determine the winning neuron, and the solution of (ii) has closed-form expression given by (4.3).

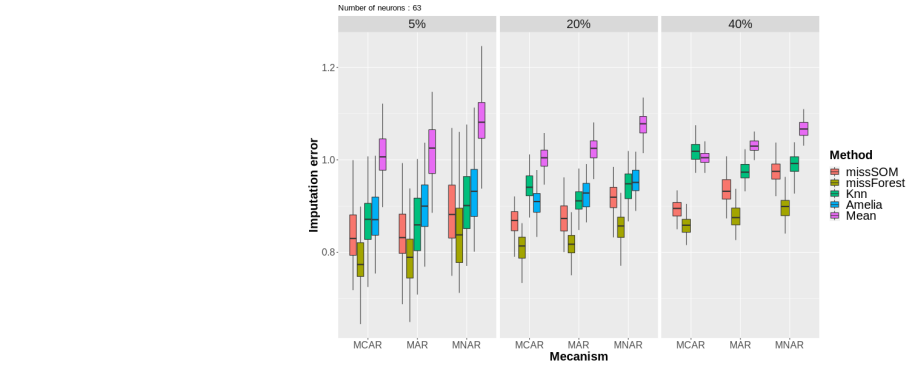
4.3.4 Numerical performance

In a numerical study we compare missSOM to other methods. The simplest approach to deal with missing data consists in simply deleting the observations that have missing values and applying the standard Kohonen algorithm for complete data to the remaining data. In the simulations this approach is shown to be far from optimal and all error rates are the worst.

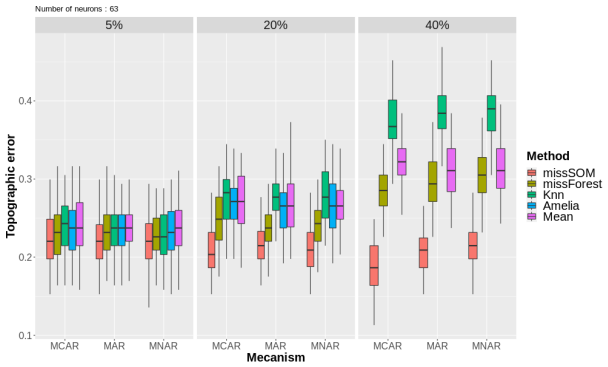
The approach proposed by Cottrell *et al.* [161] is suited for incomplete observations, but does not incorporate any estimation of missing values within the algorithm. However, imputation is easily performed once the map is learned by imputing the values of the winning prototypes. The quality of the map obtained with Cottrell's method is very close to that obtained by missSOM. However, missSOM achieves a significantly better imputation error. This highlights that it is beneficial to learn missing values simultaneously with the map, and not separately afterwards.

In comparison with other imputation methods, missSOM is doing fine. While its imputation error is slightly worse than that of missForest by [163], which uses random forests, it is better than other imputation methods like k-nearest neighbor (kNN) (see [164]), *amelia*, a model-based approach based on Gaussian mixtures model [165], or just imputing by the mean value. This is illustrated in Figure 4.2 (a) on the wines dataset from the UCI machine learning repository [166], where we generated missing entries randomly with various proportions.

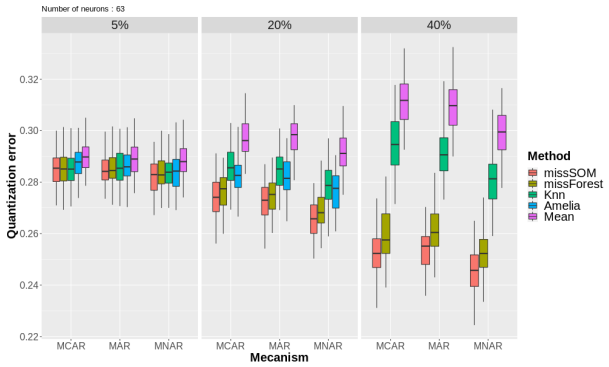
Now, with any imputation method at hand, we can first impute missing values and then apply the standard Kohonen algorithm for complete data. The maps obtained that way are all much worse than with missSOM, as can be seen by the topographic and the quantization errors in Figure 4.2 (b) and (c). Again, this shows that our simultaneous learning approach yields better results.



(a) Imputation error



(b) Topographic error



(c) Quantization error

Figure 4.2: Different errors of *missSOM* and classical imputation methods on the wines data including missing entries.

4.4 Control of the false clustering rate

When clustering data, it often occurs that data sets include ambiguous individuals that are intrinsically difficult to attribute to one cluster or another. This is the case of outliers or data points the fall in the overlap of two clusters. In applications, misclassifying individuals is potentially disastrous and should be avoided. To keep the misclassification rate small, one can opt to cluster only a part of the data, namely the datapoints with low uncertainty. The purpose of this work is the development of a method with an abstention option that comes with the guarantee that the false clustering rate (FCR) does not exceed a predefined nominal level α.

This is a work with my student Ariane Marandon and my colleagues Etienne Roquain and Nataliya Sokolovska [MRRS22].

4.4.1 State of the art

In a supervised setting, classification with an abstention option is a long-standing statistical paradigm, see [167, 168, 169, 170] among others. In this line of research, rejection or the decision not to classify an observation is accounted for by adding a term to the risk that penalizes any rejection. Recently, still in the supervised setting, [171] and [172] propose a method that controls an error among the classified items at prescribed level. These methods consist in thresholding the estimated class probabilities in a data-driven manner.

In [173] an approach similar to ours is presented, but the control of the false clustering rate is only established in the case of known model parameters. Our work goes much further and provides guarantees for the completely data-driven procedure.

The false clustering rate is closely related to the false discovery rate (FDR) in multiple testing. In fact, we can roughly view the problem of designing an abstention rule as testing problem, where, for each item i, we test whether the proposed cluster label is reliable or not. With this analogy, our selection rule is based on quantities similar to the local FDR values [174], a key quantity to build optimal FDR controlling procedures in multiple testing mixture models, see, e.g., [175, 176, 177] and [RRV22]. In particular, our final selection procedure shares similarities with the procedure introduced in [176], also named cumulative ℓ-value procedure [178]. In addition, our theoretical analysis is related to the work of [RRV22], although the nature of the algorithm developed therein is different.

4.4.2 Clustering procedures with abstention

We consider a classical finite mixture model for the sample $\mathbf{X} = (X_1, \dots, X_n)$ with latent variables $\mathbf{Z} = (Z_1, \dots, Z_n) \in [\![K]\!]^n$ and cluster probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, that is, $\mathbb{P}(Z_i = k) = \pi_k$. Conditionally on $\mathbf{Z}$,

$$\mathbf{X}|\mathbf{Z} = \bigotimes_{i \in [\![n]\!]} X_i|Z_i = \bigotimes_{i \in [\![n]\!]} f_{\phi_{Z_i}},$$

where the densities $f_{\phi_k} \in \{f_u, u \in \mathcal{U}\}$ for $k \in [\![K]\!]$ belong to a parametric family of densities on $\mathbb{R}^d$ with parameter $\boldsymbol{\phi} = (\phi_1, \dots, \phi_K)$. The number of clusters K is assumed to be known. The goal is to recover the cluster labels $\mathbf{Z}$ from the data $\mathbf{X}$.

Let $\widehat{\mathbf{Z}} = (\hat{Z}_i)_{i \in [\![n]\!]} \in [\![K]\!]^n$ be any clustering rule obtained on the data $\mathbf{X}$. In the unsupervised setting only the partition of the observations is of interest, not the labels themselves. Switching the labels of $\widehat{\mathbf{Z}}$ does not change the corresponding partition. Let $S \subset [\![n]\!]$ be a selection rule, indicating the indices of observations that are clustered. Now a clustering procedure as considered in this work is defined by a clustering rule and a selection rule, say $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$.

The classification error is defined by $\varepsilon_S(\widehat{\mathbf{Z}}, \mathbf{Z}) = \frac{1}{|S|} \sum_{i \in S} \mathbb{1}\{Z_i \neq \hat{Z}_i\}$, which depends on the order of the label. We define the label-switching invariant *false clustering rate* (FCR) as

$$\mathrm{FCR}(\mathcal{C}) = \mathbb{E}\left[\min_{\sigma \in \mathfrak{S}(K)} \mathbb{E}\left[\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) | \mathbf{X} \right] \right],$$

where $\mathfrak{S}(K)$ denotes the set of permutations on $[\![K]\!]$. The aim is to construct a procedure with a control of the FCR at some nominal level α.

Oracle procedure

In model-based approaches, the clustering is naturally based on the posterior probabilities of the labels given by

$$\ell_k(X_i) = \mathbb{P}(Z_i = k | X_i) = \frac{\pi_k f_{\phi_k}(X_i)}{\sum_{\ell \in [\![K]\!]} \pi_\ell f_{\phi_\ell}(X_i)}.$$

When all observations are clustered, that is, $S = [\![n]\!]$, the Bayes clustering, say $\widehat{\mathbf{Z}}^*$, defined by the MAP estimators of the labels, that is $\widehat{Z}_i^* = \arg\max_{k\in[\![K]\!]} \ell_k(X_i)$, has minimal FCR. However, depending on the intrinsic difficulty of the clustering problem, the FCR of the Bayes clustering can exceed the nominal level α.

The quantities $T(X_i) = 1 - \max_{k\in[\![K]\!]} \ell_k(X_i)$ indicate the uncertainty of the MAP cluster labels, and one can show that the FCR of the Bayes clustering is given by

$$\mathrm{FCR}((\widehat{\mathbf{Z}}^*, S)) = \mathbb{E}\left[\frac{1}{|S|}\sum_{i\in S} T(X_i)\right].$$

Thus, to obtain the control of the FCR, the best way consists in not selecting the most ambiguous observations that contribute the most to the FCR of the Bayes clustering, that is, those with the largest values $T(X_i)$. We consider a thresholding-based selection rule of the form $S = \{i \in [\![n]\!] : T(X_i) \leq t\}$, where threshold t is chosen such that $\sum_{i\in S} T(X_i) \leq \alpha|S|$ while maximizing $|S|$. This gives rise to the *oracle procedure*, that can be easily implemented by ordering the values $T(X_i)$.

Plug-in and bootstrap procedures

The oracle procedure cannot be used in practice since it is based on the knowledge of the true model parameter θ^*. A natural idea then is to replace θ^* by an estimator $\hat{\theta}$ giving rise to the *plug-in procedure*. Despite very favorable theoretical properties of the plug-in procedure, its FCR can exceed α in particular when the estimator $\hat{\theta}$ is too rough. Indeed, the uncertainty of $\hat{\theta}$ near θ^* is ignored by the plug-in procedure.

To fix this issue, a bootstrap approximation of the FCR of the plug-in procedure can be used and then threshold t is chosen such that bootstrapped FCR is controlled by α. Both parametric and nonparametric bootstrap can be used here. The parametric bootstrap provides good results when the $\hat{\theta}$ is an accurate estimate. Otherwise the nonparametric bootstrap approach is preferable.

4.4.3 Optimality results

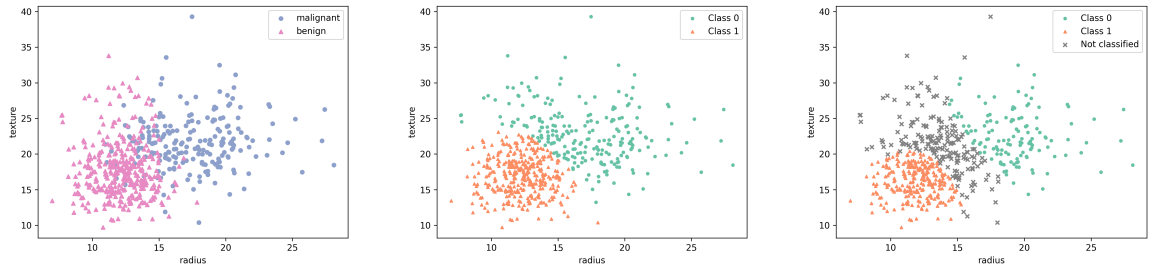
We show that the plug-in procedure $\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}}$ has (almost) optimal behavior. More precisely, its FCR, and the so-called marginal FCR, defined as a ratio of expectations instead of an expectation of a ratio, are close to α, while the mean number of selected observations is nearly optimal. We provides both consistency and convergence rates.

Theorem 12. *Under appropriate assumptions, the plug-in procedure $\widehat{\mathcal{C}}_\alpha^{PI}$ satisfies*

$$\limsup_n \mathrm{FCR}(\widehat{\mathcal{C}}_\alpha^{PI}) \leq \alpha, \quad \limsup_n \mathrm{MFCR}(\widehat{\mathcal{C}}_\alpha^{PI}) \leq \alpha,$$

and for any procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ that controls the marginal MFCR at level α, it holds

$$\liminf_n \{n^{-1}\mathbb{E}_{\theta^*}(|\widehat{S}_\alpha^{PI}|) - n^{-1}\mathbb{E}_{\theta^*}(|S|)\} \geq 0.$$



(a) Ground truth labels (b) MAP clustering (c) MAP clustering

Figure 4.3: Comparison of clusterings of the variables *radius* and *texture* in the WDBC dataset.

Moreover, there exist constants C_1 *and* C_2 *such that for any sequence* $\varepsilon_n = o(1)$ *and any* n *large enough,*

$$\max((\mathrm{FCR}(\widehat{\mathcal{C}}_\alpha^{PI}), \mathrm{MFCR}(\widehat{\mathcal{C}}_\alpha^{PI})) \leq \alpha + C_1 \left(\varepsilon_n^{1/2} + \sqrt{(\log n)/n} + \eta(\varepsilon_n, \theta^*)\right)$$
$$n^{-1}\mathbb{E}_{\theta^*}(|\widehat{S}_\alpha^{PI}|) - n^{-1}\mathbb{E}_{\theta^*}(|S|) \geq -C_2 \left(\varepsilon_n^{1/2} + \sqrt{(\log n)/n} + \eta(\varepsilon_n, \theta^*)\right),$$

for any procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ *that controls the marginal MFCR at level* α *(non asymptotically) and* $\eta(\varepsilon, \theta^*) = \mathbb{P}_{\theta^*}\left(\min_{\sigma \in \mathfrak{S}(K)} \|\hat{\theta}^\sigma - \theta^*\|_2 \geq \varepsilon\right)$*, a quantity measuring the quality of the estimator.*

The proofs employ techniques similar to those used in [RRV22] for multiple testing of pairwise hypotheses.

4.4.4 Numerical performance

In a simulation study, the parametric bootstrap procedure shows overall the more "stable" behavior: it uniformly improves the plug-in procedure across all the explored parameter ranges. In addition, it achieves an FCR and a selection frequency close to those of the oracle when the sample size n is fairly large. For more challenging cases, where the sample size is small and a strict FCR control is desired, the non-parametric bootstrap is a valuable alternative.

To conclude, we demonstrate our method on the Wisconsin Breast Cancer Diagnosis (WDBC) dataset from the UCI ML repository consisting of 30 features computed from a digitalized image of a fine needle aspirate (FNA) of a breast mass, on a total of 569 patients, of which 212 are diagnosed as benign and 357 as malignant. A mixture of Student's t-distributions is chosen to model the data [179] to account for outliers leading to overlapping clusters. As the t-distribution is less concentrated than the Gaussian, this may help with the estimation of the posterior probabilities of the cluster labels, which are typically overestimated in Gaussian mixtures.

Figure 4.3 (a) displays the first two variables of the dataset, the radius and the texture of the images, and the labels are the classes benign and malignant. For a nominal level $\alpha = 5\%$, the MAP clustering without any selection (see (b)) achieves an FCR of 14%. Our parametric bootstrap procedure in (c) controls the FCR by clustering only 70% of the data. Its FCR equals

3%, which is below the target level.

Chapter 5

Perspectives

This chapter provides perspectives and directions of research raising from the work presented in this manuscript. I currently explore some of these open questions, others are left for future investigations. Generally speaking, my current and future research shall be in line with my past research. I plan to contribute to the development of new methods for general problems in statistics and machine learning. Moreover, in collaboration with scientists from various fields, I wish to design specific statistical solutions for application-specific problems. More precisely, building on my expertise in statistical modelling and the development of efficient algorithms, I plan to contribute to the following topics.

Multiple networks

For a long time, research has focused on the analysis of single networks. Today the interest turns to the joint analysis of multiple networks, as collections of networks are available in more and more fields of application. Naturally, questions emerge on graph comparison, graph embeddings, modelling, inference, scalability of algorithms and many others. Our work on graph clustering (Section 2.4) in particular raises a number of further questions.

Consistency in the mixture model of SBMs To start with, a theoretical study of the mixture model of SBMs and its estimates is in order. In an asymptotic setting where both the number of vertices per network $n^{(m)}$ and the number of observed networks M tend to infinity, it is plausible that consistency of the maximum likelihood estimator is inherited from the consistency of the estimate in a single SBM with increasing number of nodes. However, when only the number of networks M tends to infinity, the situation is different as adding new networks to the data set may improve parameter estimates, but not directly the node clusterings. Especially in small networks, even with the knowledge of the true parameter value, there may be nodes that are inherently difficult to attribute to a block. Thus, it is expected that node clusterings are less accurate in that case and it would be interesting to quantify the extent of that loss.

Reliable graph clustering results In Section 4.4 we proposed a method for clustering vectors with an abstention option and a control of the false clustering rate. This tool may be adapted for graph clustering and used in association with the hierarchical clustering algorithm proposed in Section 2.4. This requires to analyze the the posterior probabilities that a network belongs to a given cluster in a mixture model of SBMs. The resulting procedure with a control of the false clustering rate could be used to check whether there is evidence that a network really belongs to the cluster it is assigned to or not. As such the final clustering may be improved yielding more reliable clusters. A related question, that is also interesting to study, is outlier detection, where the goal is to test if a network was generated by the fitted model or by a different distribution.

Submodels In some settings where networks with different topologies are observed, the variability may not be due to the fact that data are sampled from independent mixture components, but rather that they are subsampled from one huge network. This could be, for instance, the case in ecology, where the observer chooses to focus on a specific set of species and reports only interactions among the chosen species while others are ignored. In general, subsampling of networks induces bias and thus can be the reason why different topologies are observed in a collection of networks. Thus, to go further in the analysis of the graph clustering result of the algorithm presented in Section 2.4, there would be an interest to develop a statistical test to compare two clusters. More precisely, one would wish to test whether a cluster is just a submodel of the SBM of another cluster or if there is a substantial difference in the topology.

This question is part of the research branch that deals with subsampling and resampling of networks, which are involved and up-to-date problems frequently addressed in the literature during the last years [180, 181, 182, 183]. Indeed, it would be helpful to have a statistical model for networks subsampled from a huge network. There might be a number of identifiability issues and the development of an inference method might be challenging but very useful for practice.

Gaussian graphical models Concerning the noisy stochastic block model (NSBM) that we developed for a multiple testing task in Section 2.3, a different perspective consists in linking our approach to the inference of Gaussian graphical models (GGM). Based on recent estimates of the precision matrix [44, 184, 185], the NSBM may be adapted for modelling. This is related to the approach in [186] where a SBM is introduced for the precision matrix. An adaptive multiple testing procedure in the vein of the one for testing paired null hypotheses may be derived that comes with a power enhancement for the inference in the GGM.

Network analysis for specific applications

Ecological networks Our study of clustering foodwebs in Section 2.4.5 has raised the interest of ecologists and together we might deepen the analysis of the mangal database [67] to address more specific questions in ecology. The graph clustering and the modelling by SBMs seem to be fruitful for a better understanding of the organization and functioning of ecological networks,

which is particularly important in the light of climate change. This work shall be done in tight collaboration with environmental scientists.

Metabolic networks Since graph clustering is a relatively new research area, many specific settings remain to be explored and may require different solutions. In the Ph.D. of Ariane Marandon we study a database that contains hundreds of metabolic networks of bacteria. We work on a model that considers observed networks as perturbations of an unknown underlying network. By introducing a mixture model, we aim at performing graph clustering of the metabolic networks. The obtained clustering may help to study the impact of environmental factors on the metabolism of bacteria, such as the temperature of the environment where the bacteria lives. Another question is, for instance, whether an automated clustering is capable of uncovering the phylogeny of bacteria.

Networks in epidemiology In epidemiology of cattle diseases, a profound understanding of the animal trade network is highly valuable to control pathogen spread. With researchers at INRAE, we work on a scale-free percolation model (SFP) [187] and aim at developing a scalable estimation algorithm. At this stage, a major hindrance to use the SFP model on the French Database of Cattle Movements (FDCM), maintained by the Ministry of Agriculture, is the sparsity of the observed network that renders estimates very unstable. To build estimators that are consistent in the sparse setting, we may exploit the specific structure of the graph, which is organized in communities due to the geography of the farms. Roughly, the idea consists in considering relatively dense subgraphs, on which individual SFP models could be fitted with high accuracy. Then the fitted model should be aggregated to a unique SFP model for the entire network. The main questions here are how to perform subsampling of the graph, and how to correctly aggregate models.

Computational statistics

Theoretical foundation of the ICL maximization A rather new inference algorithm for discrete latent variable models is based on the integrated classification likelihood (ICL) introduced by [87]. In several models, like the mixture of SBMs proposed in Section 2.4, the method achieves good performances and the hierarchical clustering algorithm is very attractive as it provides a nested sequence of clusterings and comes with an automatic selection of the number of clusters. Furthermore, the algorithm has computational advantages compared to other algorithms as EM, for instance, and it is scalable to large data sets. In future work, theoretical foundations of the ICL maximization approach might be worked out. Questions as the consistency of the estimators and of the model selection should be addressed and also the limitations of the ICL approach should be studied.

Study of variational EM algorithms In the work on multiple testing in Section 2.3, where the problem is recast as a graph inference problem, we have seen that the convergence rates of the false and true discovery rates depend on the quality of the parameter estimates. To go

further in this study, a deeper understanding of the clustering obtained by the variational EM algorithm in the noisy stochastic block model is required. This task meets several recent works on various versions of the SBM, see [58, 59, 60], but the solutions proposed therein are still incomplete to make our convergence rates explicit.

Mini-batch sampling The work on mini-batch sampling in the MCMC-SAEM algorithm described in Section 2.2 raises (at least) two questions, which are also of interest in other algorithms based on mini-batch sampling. First, there is the question of the optimal mini-batch proportion α operating a trade-off between speeding up convergence and estimation accuracy. To get closer to this goal, it would be valuable to determine the limit distribution of the estimates and, in particular, to quantify the impact of the mini-batch proportion on the limit variance. In our paper [KMR20] we present some heuristic arguments in favor of asymptotic normality of the estimator and we conjecture a formula of the limit variance. These findings are supported by our simulation study, but a formal proof is still missing. Second, when given a limited computing time budget, which should be the case in most practical situations, it would be extremely useful to have a strategy of how to choose the optimal mini-batch proportion and/or the best data set as illustrated in the example on the model for handwritten digits (Section 2.2.5), to achieve the best possible results in the given time limit.

Scalable EM-type algorithms Related to mini-batch approaches, another avenue of research to speed up computation are variance reduction techniques. The works of [188] and [189] propose extensions to the EM algorithm using Stochastic Variance-Reduced Gradient (SVRG) [190] and the Stochastic Average Gradient Algorithm (SAGA) [191] techniques, respectively. More recently, in [192] the EM algorithm is combined with the so-called Stochastic Path-Integrated Differential EstimatoR (SPIDER) for smooth non-convex problems introduced by [193]. In all these approaches the variance is reduced by introducing a control variate, and they differ in the way to construct the control variate. Our ambition is to propose similar methods for various classes of EM algorithms and provide theoretical and numerical evidence for the improvement. Moreover, motivated by research in plant improvement at INRAE, our goal is an efficient implementation of such algorithms for statistical models with high-dimensional latent variables.

Deep learning

Generative models for chemical structures A new trend in material science is to rely on artificial intelligence to discover new valuable chemical compounds via generative models [194]. For the purpose of hydrogen storage, there is a particular interest in new stable crystal structures with specified properties. The most recent generative models that produce impressive results on images rely on diffusion models and denoising score matching with Langevin dynamics [195, 196]. The goal of the Ph.D. thesis of Arsen Sultanov is the adaptation of those models to the problem of crystal generation by incorporating a large number of constraints like periodicity,

rotation invariance of atomic positions and symmetry groups in the model. This is a work in collaboration with the Institut de Chimie et des Matériaux Paris-Est.

Graph neural networks Graph neural networks is the umbrella term for various neural networks that operate on graph structured data. They have grown rapidly in scope and popularity in recent years and are appropriate when some graph signal is observed, see the survey in [73]. However, in the absence of graph signal suitable methods are still lacking. This is contrary to statistics, where the standard setting is the one where the only available information is the graph itself. Using the statistical background, new deep learning models that solely analyze the graph topology may be proposed.

Bibliography

[1] E. D. Kolaczyk, *Statistical Analysis of Network Data Methods and Models*. New York, NY: Springer, 2009.

[2] M. Salter-Townshend, A. White, I. Gollini, and T. B. Murphy, "Review of statistical network analysis: models, algorithms, and software," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 4, pp. 243–264, 2012.

[3] B. Klimt and Y. Yang, "The Enron Corpus: A new dataset for email classification research," in *Machine Learning: ECML 2004* (J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, eds.), vol. 3201 of *Lecture Notes in Computer Science*, pp. 217–226, Springer Berlin Heidelberg, 2004.

[4] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, and et al., "High-resolution measurements of face-to-face contact patterns in a primary school," *PLoS ONE*, vol. 6, no. 8, p. e23176, 2011.

[5] I. Gollini and T. B. Murphy, "Joint modeling of multiple network views," *Journal of Computational and Graphical Statistics*, vol. 25, no. 1, pp. 246–265, 2016.

[6] P. Holme, "Modern temporal network theory: a colloquium," *Eur. Phys. J. B*, vol. 88, no. 9, p. 234, 2015.

[7] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding, *Statistical models based on counting processes*. Springer Series in Statistics, Springer-Verlag, New York, 1993.

[8] C. T. Butts, "A relational event framework for social action," *Sociol. Methodol.*, vol. 38, no. 1, pp. 155–200, 2008.

[9] D. Q. Vu, D. Hunter, P. Smyth, and A. U. Asuncion, "Continuous-time regression models for longitudinal networks," in *Adv Neural Inf Process Syst 24* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds.), pp. 2492–2500, Curran Associates, Inc., 2011.

[10] P. O. Perry and P. J. Wolfe, "Point process modelling for directed interaction networks," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 75, no. 5, pp. 821–849, 2013.

[11] M. A. Kramer, U. T. Eden, S. S. Cash, and E. D. Kolaczyk, "Network inference with confidence from multivariate time series," *Physical Review E*, vol. 79, no. 6, 2009.

[12] C. Matias and S. Robin, "Modeling heterogeneity in random graphs through latent space models: a selective review," *Esaim Proc. & Surveys*, vol. 47, pp. 55–74, 2014.

[13] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, "Detecting communities and their evolutions in dynamic social networks—a Bayesian approach," *Mach. Learn.*, vol. 82, no. 2, pp. 157–189, 2011.

[14] K. Xu and A. Hero, "Dynamic stochastic blockmodels for time-evolving social networks," *IEEE J. Sel. Topics Signal Process.*, vol. 8, pp. 552–562, Aug 2014.

[15] M. Corneli, P. Latouche, and F. Rossi, "Exact ICL maximization in a non-stationary temporal extension of the stochastic block model for dynamic networks," *Neurocomputing*, vol. 192, pp. 81 – 91, 2016.

[16] C. Matias and V. Miele, "Statistical clustering of temporal networks through a dynamic stochastic block model," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 79, no. 4, pp. 1119–1141, 2017.

[17] D. Böhning, "A review of reliable maximum likelihood algorithms for semiparametric mixture models," *J. Stat. Plan. Inference*, vol. 47, no. 1–2, pp. 5 – 28, 1995.

[18] L. Bordes, D. Chauveau, and P. Vandekerkhove, "A stochastic EM algorithm for a semiparametric mixture model," *Comput. Stat. Data Anal.*, vol. 51, no. 11, pp. 5429 – 5443, 2007.

[19] S. Robin, A. Bar-Hen, J.-J. Daudin, and L. Pierre, "A semi-parametric approach for mixture models: Application to local false discovery rate estimation," *Comput. Stat. Data Anal.*, vol. 51, no. 12, pp. 5483 –

5493, 2007.

[20] J. Dannemann, "Semiparametric Hidden Markov models," *J. Comput. Graph. Statist.*, vol. 21, no. 3, pp. 677–692, 2012.

[21] J.-J. Daudin, F. Picard, and S. Robin, "A mixture model for random graphs," *Statist. Comput.*, vol. 18, no. 2, pp. 173–183, 2008.

[22] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[23] P. Reynaud-Bouret, "Penalized projection estimators of the Aalen multiplicative intensity," *Bernoulli*, vol. 12, no. 4, pp. 633–661, 2006.

[24] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Trans. Pattern Anal. Machine Intel.*, vol. 22, no. 7, pp. 719–725, 2000.

[25] Transport for London, "Cycle hire usage data 2012 - 2015." http://cycling.data.tfl.gov.uk/, 2016.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[27] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models* (M. I. Jordan, ed.), pp. 355–368, Cambridge, MA, USA: MIT Press, 1999.

[28] P. Liang and D. Klein, "Online EM for unsupervised models," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pp. 611–619, Association for Computational Linguistics, 2009.

[29] B. Karimi, M. Lavielle, and É. Moulines, "On the Convergence Properties of the Mini-Batch EM and MCEM Algorithms." preprint `https://hal.inria.fr/hal-02334485`, 2019.

[30] H. Nguyen, F. Forbes, and G. McLachlan, "Mini-batch learning of exponential family finite mixture models," *Statist. Comput.*, 2020.

[31] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Statist. Soc. Ser. B*, vol. 2, no. 46, pp. 257–267, 1984.

[32] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 2, pp. 425–437, 01 1995.

[33] O. Cappé and E. Moulines, "On-line expectation-maximization algorithm for latent data models," *J. Roy. Statist. Soc. Ser. B*, vol. 71, no. 3, pp. 593–613, 2009.

[34] O. Cappé, "Online EM algorithm for hidden Markov models," *Journal Computational and Graphical Statistics*, vol. 20, no. 3, pp. 728–749, 2011.

[35] E. Kuhn and M. Lavielle, "Coupling a stochastic approximation version of EM with an MCMC procedure," *ESAIM: P&S*, vol. 8, pp. 115–131, 2004.

[36] C. P. Robert and G. Casella, *Monte Carlo statistical methods*. Springer Texts in Statistics, Springer-Verlag, New York, second ed., 2004.

[37] S. Allassonnière, Y. Amit, and A. Trouvé, "Toward a coherent statistical framework for dense deformable template estimation," *J. Roy. Statist. Soc. Ser. B*, vol. 69, pp. 3–29, 2007.

[38] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994.

[39] W. Sun and T. T. Cai, "Large-scale multiple testing under dependence," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 71, no. 2, pp. 393–424, 2009.

[40] N. Eagle and A. (Sandy) Pentland, "Reality mining: Sensing complex social systems," *Journal of Personal and Ubiquitous Computing*, vol. 10, p. 255–268, Mar. 2006.

[41] D. Lusseau, "Evidence for social role in a dolphin social network," *Evolutionary Ecology*, vol. 21, p. 357–366, 2007.

[42] U. Nations. `https://www.un.org/en/development/desa/population/migration/data`, 2019.

[43] M. Drton, M. D. Perlman, *et al.*, "Multiple testing and error control in gaussian graphical model selection," *Statistical Science*, vol. 22, no. 3, pp. 430–449, 2007.

[44] W. Liu, "Gaussian graphical model estimation with false discovery rate control," *Ann. Statist.*, vol. 41, no. 6, pp. 2948–2978, 2013.

[45] T. T. Cai and W. Liu, "Large-scale multiple testing of correlations," *J. Amer. Statist. Assoc.*, vol. 111, no. 513, pp. 229–240, 2016.

[46] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical Bayes analysis of a microarray experiment," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1151–1160, 2001.

[47] B. Efron, "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 96–104, 2004.

[48] W. Sun and T. T. Cai, "Oracle and adaptive compound decision rules for false discovery rate control," *J. Amer. Statist. Assoc.*, vol. 102, no. 479, pp. 901–912, 2007.

[49] T. T. Cai and W. Sun, "Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks," *J. Amer. Statist. Assoc.*, vol. 104, no. 488, pp. 1467–1481, 2009.

[50] J. Liu, C. Zhang, and D. Page, "Multiple testing under dependence via graphical models," *Ann. Appl. Stat.*, vol. 10, no. 3, pp. 1699–1724, 2016.

[51] J. Chang, E. Kolaczyk, and Q. Yao, "Estimation of subgraph densities in noisy networks," *Journal of the American Statistical Association*, pp. 1–14, 2020.

[52] M. Newman, "Network structure from rich but noisy data," *Nature Physics*, vol. 14, pp. 542–545, 2017.

[53] M. Newman, "Estimating network structure from unreliable measurements," *Physical Review E*, vol. 98, 2018.

[54] C. M. Le, K. Levin, and E. Levina, "Estimating a network from multiple noisy realizations," *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 4697 – 4740, 2018.

[55] T. P. Peixoto, "Reconstructing networks with unknown and heterogeneous errors," *Phys. Rev. X*, vol. 8, p. 041011, Oct 2018.

[56] J.-G. Young, G. T. Cantwell, and M. Newman, "Robust bayesian inference of network structure from unreliable data," *J. Complex Networks*, vol. 8, 2021.

[57] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[58] A. Celisse, J.-J. Daudin, and L. Pierre, "Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model," *Electron. J. Statist.*, vol. 6, pp. 1847–1899, 2012.

[59] P. Bickel, D. Choi, X. Chang, and H. Zhang, "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels," *Ann. Statist.*, vol. 41, pp. 1922–1943, 08 2013.

[60] V. Brault, C. Keribin, and M. Mariadassou, "Consistency and asymptotic normality of latent blocks model estimators," *Electronic Journal of Statistics*, vol. 14, no. 1, pp. 1234–1268, 2020.

[61] T. T. Cai, W. Sun, and W. Wang, "Covariate-assisted ranking and screening for large-scale two-sample inference," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 81, no. 2, pp. 187–234, 2019.

[62] T. Schweder and E. Spjøtvoll, "Plots of P-values to evaluate many tests simultaneously," *Biometrika*, vol. 69, no. 3, pp. 493–502, 1982.

[63] J. D. Storey, "A direct approach to false discovery rates," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 64, no. 3, pp. 479–498, 2002.

[64] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. V. den Broeck, "What's in a crowd? analysis of face-to-face behavioral networks," *Journal of Theoretical Biology*, vol. 271, no. 1, pp. 166–180, 2011.

[65] C. Donnat and S. Holmes, "Tracking network dynamics: A survey using graph distances," *The Annals of Applied Statistics*, vol. 12, no. 2, pp. 971 – 1012, 2018.

[66] A. Weber-Zendrera, N. Sokolovska, and H. A. Soula, "Functional prediction of environmental variables using metabolic networks," *Scientific Reports*, vol. 11, p. 12192, 2021.

[67] T. Poisot, B. Baiser, J. A. Dunne, S. Kéfi, F. c. Massol, N. Mouquet, T. N. Romanuk, D. B. Stouffer, S. A. Wood, and D. Gravel, "mangal – making ecological network analysis simple," *Ecography*, vol. 39, no. 4, pp. 384–390, 2016.

[68] T. Gärtner, "A survey of kernels for structured data," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, pp. 49–58, 2003.

[69] N. Shervashidze, S. Vishwanathan, T. Petri, K. Mehlhorn, and K. Borgwardt, "Efficient graphlet kernels

for large graph comparison," in *JMLR Workshop and Conference Proceedings Volume 5: AISTATS 2009*, (Cambridge, MA, USA), pp. 488–495, Max-Planck-Gesellschaft, MIT Press, Apr. 2009.

[70] Y. Shimada, Y. Hirata, T. Ikeguchi1, and K. Aihara, "A survey of kernels for structured data," *Scientific Reports*, vol. 6, p. 34944, 2016.

[71] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *IEEE Data Engineering Bulletin*, vol. 40, no. 3, pp. 52–74, 2017.

[72] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *International Conference on Learning Representations*, 2019.

[73] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[74] L. le Gorrec, P. A. Knight, and A. Caen, "Learning network embeddings using small graphlets," *Social Network Analysis and Mining*, vol. 12, no. 20, 2022.

[75] C. E. Ginestet, J. Li, P. Balachandran, S. Rosenberg, and E. D. Kolaczyk, "Hypothesis testing for network data in functional neuroimaging," *The Annals of Applied Statistics*, vol. 11, no. 2, pp. 725–750, 2017.

[76] H. Zanghi, S. Volant, and C. Ambroise, "Clustering based on random graph model embedding vertex features," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 830–836, 2010.

[77] I. C. Gormley, T. B. Murphy, and A. E. Raftery, "Model-based clustering," *Annual Review of Statistics and Its Application*, vol. 10, no. 1, 2023.

[78] N. Stanley, S. Shai, D. Taylor, and P. J. Mucha, "Clustering network layers with the strata multilayer stochastic block model," *IEEE Transactions on Network Science and Engineering*, vol. 3, pp. 95–105, apr 2016.

[79] M. Sabanayagam, L. C. Vankadara, and D. Ghoshdastidar, "Graphon based clustering and testing of networks: Algorithms and theory," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[80] S. S. Mukherjee, P. Sarkar, and L. Lin, "On clustering network-valued data," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[81] M. Signorelli and E. C. Wit, "Model-based clustering for populations of networks," *Statistical Modelling*, vol. 20, no. 1, pp. 9–29, 2019.

[82] A. Mantziou, S. Lunagomez, and R. Mitra, "Bayesian model-based clustering for multiple network data," 2021. `https://arxiv.org/abs/2107.03431`.

[83] J.-G. Young, A. Kirkley, and M. E. J. Newman, "Clustering of heterogeneous populations of networks," *Physical Review E*, vol. 105, jan 2022.

[84] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. Wiley series in probability and statistics, Wiley, 2. ed ed., 2008.

[85] J. Liu, *Monte Carlo strategies in scientific computing*. New York, Berlin, Heidelberg: Springer Verlag, 2008.

[86] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[87] E. Côme and P. Latouche, "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood," *Statistical Modelling*, vol. 15, no. 6, pp. 564–589, 2015.

[88] L. Lovász and B. Szegedy, "Limits of dense graph sequences," *Journal of Combinatorial Theory, Series B*, vol. 96, no. 6, pp. 933–957, 2006.

[89] P. J. Bickel and A. Chen, "A nonparametric view of network models and newman–girvan and other modularities," *Proceedings of the National Academy of Sciences*, vol. 106, no. 50, pp. 21068–21073, 2009.

[90] Y. Vardi, "Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation," *Biometrika*, vol. 76, pp. 751–761, 1989.

[91] P. Kvam, "Length bias in the measurements of carbon nanotubes," *Technometrics,*, vol. 50, no. 4, pp. 462–467, 2008.

[92] F. Balabdaoui and J. A. Wellner, "Estimation of a k-monotone density: limit distribution theory and the spline connection," *Annals of Statistics*, vol. 35, no. 6, pp. 2536–2564, 2007.

[93] C.-H. Zhang, "Fourier methods for estimating mixing densities and distributions," *Annals of Statistics*,

vol. 18, pp. 806–831, 1990.

[94] C. Goutis, "Nonparametric estimation of a mixing density via the kernel method," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1445–1450, 1997.

[95] M. Asgharian, M. Carone, and V. Fakoor, "Large-sample study of the kernel density estimators under multiplicative censoring," *Annals of Statistics*, vol. 40, no. 1, pp. 159–187, 2012.

[96] J. Fan, "On the optimal rates of convergence for nonparametric deconvolution problems," *Ann. Statist.*, vol. 19, no. 3, pp. 1257–1272, 1991.

[97] N. W. Hengartner, "Adaptive demixing in poisson mixture models," *Annals of Statistics*, vol. 25, no. 3, pp. 917–928, 1997.

[98] F. Roueff and T. Ryden, "Nonparametric estimation of mixing densities for discrete distributions," *Annals of Statistics*, vol. 33, pp. 2066–2108, 2005.

[99] C.-H. Zhang, "On estimating mixing densities in discrete exponential family models," *Annals of Statistics*, vol. 23, pp. 929–945, 1995.

[100] F. Comte and V. Genon-Catalot, "Adaptive Laguerre density estimation for mixed Poisson models," *Electronic Journal of Statistics*, vol. 9, pp. 1112–1148, 2015.

[101] D. Belomestny and J. Schoenmakers, "Statistical Skorohod embedding problem and its generalizations," tech. rep., Arxiv, 2014.

[102] Z. Ditzian and V. Totik, *Moduli of Smoothness*. Springer Series in Computational Mathematics, Springer-Verlag, 1987.

[103] P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot, "Adaptive density estimation: a curse of support?," *Journal of Statistical Planning and Inference*, vol. 141, pp. 115–139, 2011.

[104] G. Gayraud, "Estimation of functionals of density support," *Mathematical Methods of Statistics*, vol. 6, pp. 26–47, 1997.

[105] A. Juditsky and S. Lambert-Lacroix, "On minimax density estimation on r," *Bernoulli*, vol. 10, no. 2, pp. 1877–220, 2004.

[106] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*. New York: Academic/Plenum, 1999.

[107] B. Valeur, *Molecular Fluorescence*. Weinheim: Wiley-VCH, 2002.

[108] D. V. O'Connor and D. Phillips, *Time-correlated single photon counting*. London: Academic Press, 1984.

[109] A. Tsodikov, "A generalized self-consistency approach," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 65, no. 3, pp. 759–774, 2003.

[110] E. Brunel, F. Comte, and A. Guilloux, "Nonparametric density estimation in presence of bias and censoring," *Test*, vol. 18, pp. 166–194, 2005.

[111] J. Fan, "On the optimal rates of convergence for nonparametric deconvolution problems," *Ann. Statist.*, vol. 19, no. 3, pp. 1257–1272, 1991.

[112] M. Pensky and B. Vidakovic, "Adaptive wavelet estimator for nonparametric density deconvolution," *Ann. Statist.*, vol. 27, no. 6, pp. 2033–2053, 1999.

[113] P. J. Diggle and P. Hall, "A fourier approach to nonparametric deconvolution of a density estimate," *J. Roy. Statist. Soc. Ser. B*, vol. 55, no. 2, pp. 523–531, 1993.

[114] F. Comte, Y. Rozenholc, and M.-L. Taupin, "Penalized contrast estimator for adaptive density deconvolution," *Canadian Journal of Statistics*, vol. 34, pp. 431–452, 2006.

[115] R. D. Gill, Y. Vardi, and J. A. Wellner, "Large sample theory of empirical distributions in biased sampling models," *Ann. Statist.*, vol. 16, no. 3, pp. 1069–1112, 1988.

[116] C. O. Wu and A. Q. Mao, "Minimax kernels for density estimation with biased data," *Ann. Inst. Statist. Math.*, vol. 48, no. 3, pp. 451–467, 1996.

[117] C. O. Wu, "A cross-validation bandwidth choice for kernel density estimates with selection biased data," *J. Multivariate Anal.*, vol. 61, no. 1, pp. 38–60, 1997.

[118] S. Efromovich, "Distribution estimation for biased data," *J. Statist. Plann. Inference*, vol. 124, no. 1, pp. 1–43, 2004.

[119] H. El Barmi and J. S. Simonoff, "Transformation-based density estimation for weighted distributions," *J. Nonparametr. Statist.*, vol. 12, no. 6, pp. 861–878, 2000.

[120] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probability Theory*

and Related Fields, vol. 113, pp. 301–413, 1999.

[121] S. Efromovich, "Density estimation for biased data," *Ann. Statist.*, vol. 32, no. 3, pp. 1137–1161, 2004.

[122] A. Goldenshluger and O. Lepski, "Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality," *Ann. Statist.*, vol. 39, no. 3, pp. 1608–1632, 2011.

[123] A. Delaigle and I. Gijbels, "Practical bandwidth selection in deconvolution kernel density estimation," *Comput. Statist. Data Anal.*, vol. 45, no. 2, pp. 249–267, 2004.

[124] F. Comte and C. Lacour, "Data-driven density estimation in the presence of additive noise with unknown distribution," *Journal of the Royal Statistical Society B*, vol. 73, pp. 601–627, 2011.

[125] F. Navarro, C. Chesneau, and J. Fadili, "On adaptive wavelet estimation of a class of weighted densities," *Communications in Statistics - Simulation and Computation*, vol. 44, no. 8, pp. 2137–2150, 2015.

[126] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *Annals of Statistics*, vol. 24, pp. 508–539, 1996.

[127] C. Lacour and P. Massart, "Minimal penalty for goldenshluger–lepski method," *Stochastic Processes and their Applications*, vol. 126, no. 12, pp. 3774–3789, 2016. In Memoriam: Evarist Giné.

[128] C. Butucea, "Two adaptive rates of convergence in pointwise density estimation," *Math. Methods Statist.*, vol. 9, no. 1, pp. 39–64, 2000.

[129] Y. Zhang, Z. Ye, and C. Liu, "An efficient DOA estimation method in multipath environment," *Signal Processing*, vol. 90, no. 2, pp. 707–713, 2010.

[130] M. C. Vanderveen, C. B. Papadias, and A. Paulraj, "Joint angle and delay estimation (jade) for multipath signals arriving at an antenna array," *IEEE Communications Letters*, vol. 1, no. 1, pp. 12–14, 1997.

[131] C.-W. Ma and C.-C. Teng, "Detection of coherent signals using weighted subspace smoothing," *IEEE Transactions on Antennas and Propagation*, vol. 44, no. 2, pp. 179–187, 1996.

[132] T. J. Shan, M. Wax, and T. Kailath, "On spatial smoothing of estimation of coherent signals," *IEEE on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp. 806–811, 1985.

[133] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[134] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *27th Asilomar Conference on Signals, Systems and Computation*, 1993.

[135] S. Mallat, G. Davis, and Z. Zhang, "Adaptive time-frequency decompositions," *SPIE Journal of Optical Engineering*, vol. 33, pp. 2183–2191, 1994.

[136] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[137] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 91–108, 2005.

[138] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[139] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, vol. 6A, pp. 3468–3497, 2009.

[140] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, (New York, NY, USA), pp. 433–440, ACM, 2009.

[141] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, no. 84, pp. 2777–2824, 2011.

[142] A. Lozano, G. Swirszcz, and N. Abe, "Grouped orthogonal matching pursuit for variable selection and prediction," in *Advances in Neural Information Processing Systems 22*, 2009.

[143] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *The Annals of Statistics*, vol. 23, pp. 73–102, 02 1995.

[144] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecomm. ETT*, vol. 10, pp. 585–596, Nov. 1999.

[145] D. Landgrebe, "Hyperspectral image data analysis as a high dimensional signal processing problem," *Special*

Issue of the IEEE Signal Processing Magazine, vol. 19, pp. 17–28, 2002.

[146] S. Kritchman and B. Nadler, "Determining the number of components in a factor model from limited noisy data," *Chemometrics and Intelligent Laboratory Systems*, vol. 94, no. 1, pp. 19 – 32, 2008.

[147] C. F. Beckmann and S. M. Smith, "Probabilistic independent component analysis for functional magnetic resonance imaging," *IEEE Trans. Med. Imaging*, vol. 23, no. 2, pp. 137–152, 2004.

[148] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, pp. 904–909, Aug. 2006.

[149] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud, "Random matrix theory and financial correlations," *International Journal of Theoretical and Applied Finance*, vol. 3, no. 3, pp. 391–397, 2000.

[150] R. Cangelosi and A. Goriely, "Component retention in principal component analysis with application to cDNA microarray data," *Biology Direct*, vol. 2, no. 2, 2007.

[151] M. Ringnér, "What is principal component analysis?," *Nature Biotechnology*, vol. 26, no. 3, 2008.

[152] D. Passemier and J. F. Yao, "On determining the number of spikes in a high-dimensional spiked population model," *Random Matrices : Theory and Applications*, vol. 1, 2012.

[153] A. Halimi, P. Honeine, M. Kharouf, C. Richard, and J.-Y. Tourneret, "Estimating the Intrinsic Dimension of Hyperspectral Images Using a Noise-Whitened Eigengap Approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 3811–3821, 2016.

[154] V. A. Marcenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math. USSR-Sbornik*, vol. 1, pp. 457–486, 1967.

[155] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, pp. 1382–1408, 2006.

[156] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *AISTATS*, 2010.

[157] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[158] M. Hein and T. Buehler, "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA," in *NIPS*, 2010.

[159] X. Mestre, "Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5113–5129, 2008.

[160] T. Kohonen, *Self-organizing maps*. Berlin: Springer, 1995.

[161] M. Cottrell and L. P., "Missing values : processing with the Kohonen algorithm," in *ASMDA*, pp. 489–496, 2005.

[162] H. Ritter, T. Martinetz, and K. Schulten, *Neural Computation and Self-Organizing Maps: An Introduction*. USA: Addison-Wesley Longman Publishing Co., Inc., 1992.

[163] D. Stekhoven and P. Buhlmann, "Missforest - non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2011.

[164] A. Kowarik and M. Templ, "Imputation with the R package VIM," *Journal of Statistical Software*, vol. 74, no. 7, pp. 1–16, 2016.

[165] J. Honaker, G. King, and M. Blackwell, "Amelia II: A program for missing data," *Journal of Statistical Software*, vol. 45, no. 7, pp. 1–47, 2011.

[166] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[167] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.

[168] R. Herbei and M. H. Wegkamp, "Classification with reject option," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, vol. 34, no. 4, pp. 709–721, 2006.

[169] P. L. Bartlett and M. H. Wegkamp, "Classification with a reject option using a hinge loss," *Journal of Machine Learning Research*, vol. 9, no. 59, pp. 1823–1840, 2008.

[170] M. Wegkamp and M. Yuan, "Support vector machines with a reject option," *Bernoulli*, vol. 17, no. 4, pp. 1368–1385, 2011.

[171] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Proceedings of the 31st*

International Conference on Neural Information Processing Systems, NIPS'17, (Red Hook, NY, USA), p. 4885–4894, Curran Associates Inc., 2017.

[172] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei, "Learn then test: Calibrating predictive algorithms to achieve risk control," *CoRR*, vol. abs/2110.01052, 2021.

[173] T. Mary-Huard, V. Perduca, M.-L. Martin-Magniette, and G. Blanchard, "Error rate control for classification rules in multiclass mixture models," *The International Journal of Biostatistics*, vol. 18, no. 2, pp. 381–396, 2022.

[174] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1151–1160, 2001.

[175] J. D. Storey, "The positive false discovery rate: a bayesian interpretation and the q-value," *The Annals of Statistics*, vol. 31, no. 6, pp. 2013–2035, 2003.

[176] W. Sun and T. T. Cai, "Oracle and adaptive compound decision rules for false discovery rate control," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 901–912, 2007.

[177] T. Cai, W. Sun, and W. Wang, "Covariate-assisted ranking and screening for large-scale two-sample inference," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 81, no. 2, pp. 187–234, 2019.

[178] K. Abraham, I. Castillo, and É. Roquain, "Empirical Bayes cumulative ℓ-value multiple testing procedure for sparse sequences," *Electronic Journal of Statistics*, vol. 16, no. 1, pp. 2033 – 2081, 2022.

[179] D. Peel and G. J. McLachlan, "Robust mixture modelling using the t distribution," *Statistics and Computing*, vol. 10, no. 4, pp. 339–348, 2000.

[180] W. Ali, A. E. Wegner, R. E. Gaunt, C. M. Deane, and G. Reinert, "Comparison of large networks with sub-sampling strategies," *Scientific Reports*, vol. 6, no. 1, p. 28955, 2016.

[181] K. Levin and E. Levina, "Bootstrapping networks with latent space structure," 2019. preprint `https://arxiv.org/abs/1907.10821`.

[182] "Subsampling sparse graphons under minimal assumptions," 2022.

[183] J. Chang, E. D. Kolaczyk, and Q. Yao, "Estimation of subgraph densities in noisy networks," *Journal of the American Statistical Association*, vol. 117, no. 537, pp. 361–374, 2022.

[184] Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou, "Asymptotic normality and optimalities in estimation of large gaussian graphical models," *The Annals of Statistics*, vol. 43, no. 3, pp. 991–1026, 2015.

[185] J. Jankova and S. van de Geer, "Inference in high-dimensional graphical models," 2018.

[186] C. Ambroise, J. Chiquet, and C. Matias, "Inferring sparse Gaussian graphical models with latent structure," *Electronic Journal of Statistics*, vol. 3, pp. 205 – 238, 2009.

[187] M. Deijfen, R. van der Hofstad, and G. Hooghiemstra, "Scale-free percolation," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 49, no. 3, pp. 817 – 838, 2013.

[188] J. Chen, J. Zhu, Y. W. Teh, and T. Zhang, "Stochastic expectation maximization with variance reduction," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, 2018.

[189] B. Karimi, H.-T. Wai, E. Moulines, and M. Lavielle, "On the global convergence of (fast) incremental Expectation Maximization methods," in *Advances in Neural Information Processing Systems* (F. d'Alché Buc, E. Fox, and R. Garnett, eds.), vol. 32, p. 2837–2847, 2019.

[190] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, 2013.

[191] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, 2014.

[192] G. Fort, E. Moulines, and H.-T. Wai, "A Stochastic Path-Integrated Differential EstimatoR Expectation Maximization Algorithm," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[193] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, 2018.

[194] A. S. Fuhr and B. G. Sumpter, "Deep generative models for materials discovery and machine learning-accelerated innovation," *Frontiers in Materials*, 2022.

[195] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[196] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

Publications by Tabea Rebafka

[CR12] Fabienne Comte and Tabea Rebafka. Adaptive Density Estimation in the Pile-up Model Involving Measurement Errors. *Electronic Journal of Statistics*, 6:2002–2037, 2012.

[CR16] Fabienne Comte and Tabea Rebafka. Nonparametric weighted estimators for biased data. *Journal of Statistical Planning and Inference*, 174:104–128, 2016.

[CR17] Ismael Castillo and Tabea Rebafka. Discussion on 'Sparse graphs using exchangeable random measures' by F. Caron and E. B. Fox. *Journal of the Royal Statistical Society, Series B*, 79:1295–1366, 2017.

[KMR20] Estelle Kuhn, Catherine Matias, and Tabea Rebafka. Properties of the stochastic approximation EM algorithm with mini-batch sampling. *Statistics and Computing*, 30:1725–1739, 2020.

[KRS18] Malika Kharouf, Tabea Rebafka, and Nataliya Sokolovska. Consistent spectral methods for dimensionality reduction. In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, Proceedings of Machine Learning Research. EURASIP, 03–07 Sep 2018.

[MRRS22] Ariane Marandon, Tabea Rebafka, Etienne Roquain, and Nataliya Sokolovska. False clustering rate control in mixture models, 2022. arXiv:2203.02597, Submitted.

[MRV18] Catherine Matias, Tabea Rebafka, and Fanny Villers. A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680, 06 2018.

[RCF07] Tabea Rebafka, Stéphan Clémençon, and Max Feinberg. Bootstrap-based tolerance intervals for application to method validation. *Chemometrics and Intelligent Laboratory Systems*, 89:69–81, 2007.

[RDR22] Sara Rejeb, Catherine Duveau, and Tabea Rebafka. Self-organizing maps for exploration of partially observed data and imputation of missing values. *Chemometrics and Intelligent Laboratory Systems*, 231:104653, 2022.

[Reb22] Tabea Rebafka. Model-based graph clustering of a collection of networks using an agglomerative algorithm, 2022. arxiv.2211.02314, Preprint.

[RLLC11a] Tabea Rebafka, Céline Lévy-Leduc, and Maurcie Charbit. Regularization methods for intercepted radar signals. In *2011 IEEE RadarCon (RADAR)*, pages 393–396, 2011.

[RLLC11b] Tabea Rebafka, Céline Lévy-Leduc, and Maurice Charbit. OMP-type algorithm with structured sparsity patterns for multipath radar signals, 2011. arXiv:1103.5158, Technical report.

[RR15] Tabea Rebafka and François Roueff. Nonparametric estimation of the mixing density using polynomials. *Mathematical Methods of Statistics*, 24(3):200–224, 2015.

[RRS09] Tabea Rebafka, François Roueff, and Antoine Souloumiac. Procédé d'estimation des paramètres de la distribution des temps de réponse de particules d'un système, appliqué notamment aux mesures de fluorescence. Brevet numéro 09 00524, 2009.

[RRS10] Tabea Rebafka, François Roueff, and Antoine Souloumiac. A corrected likelihood approach for the pile-up model with application to fluorescence lifetime measurements using exponential mixtures. *The International Journal of Biostatistics*, 6(1), 2010.

[RRS11] Tabea Rebafka, François Roueff, and Antoine Souloumiac. Information bounds and MCMC parameter estimation for the pile-up model. *Journal of Statistical Planning and Inference*, 141(1):1–16, 2011.

[RRV22] Tabea Rebafka, Étienne Roquain, and Fanny Villers. Powerful multiple testing of paired null hypotheses using a latent graph model. *Electronic Journal of Statistics*, 16(1):2796 – 2858, 2022.

Software by Tabea Rebafka

[GMRV18] Daphné Giorgi, Catherine Matias, Tabea Rebafka, and Fanny Villers. *ppsbm: Clustering in Longitudinal Networks*, 2018. R package version 0.2.2.

[MKR20] Catherine Matias, Estelle Kuhn, and Tabea Rebafka. *R and Matlab code for the article Properties of the Stochastic Approximation EM Algorithm with Mini-batch Sampling*, 2020.

[MRV18] Catherine Matias, Tabea Rebafka, and Fanny Villers. *R code for the analysis of three datasets with the variational EM-algorithm in the Poisson process stochastic block model*, 2018.

[RDR22] Sara Rejeb, Catherine Duveau, and Tabea Rebafka. *missSOM: Self-Organizing Maps with Built-in Missing Data Imputation*, 2022. R package version 1.0.1.

[Reb23] Tabea Rebafka. *graphclust: Hierarchical Graph Clustering for a Collection of Networks*, 2023. R package version 1.0.1.

[RRV20] Tabea Rebafka, Etienne Roquain, and Fanny Villers. *R code for reproducibility of results of the article Graph inference with clustering and false discovery rate control*, 2020.

[RV20] Tabea Rebafka and Fanny Villers. *noisySBM: Noisy Stochastic Block Mode: Graph Inference by Multiple Testing*, 2020. R package version 0.1.4.